DOCUMENT RESUME

ED 248 605 EA 017 143

AUTHOR	Kennedy, Mary M.
TITLE	Working Knowledge and Other Essays.
INSTITUTION	Huron Inst., Cambridge, Mass.
SPONS AGENCY	National Inst. of Education (ED), Washington, DC.
PUB DATE	Sep 82
NOTE	236p.
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC10 Plus Postage.
DESCRIPTORS	Concept Formation; Content Analysis; *Decision Making; Elementary Secondary Education; *Evaluation Utilization; *Evaluators; Interviews; *Organizational Communication; Organizational Theories; Problem Solving; *School Administration; Social Science Research; Teacher Administrator Relationship; Teacher Evaluation; *Test Interpretation

ABSTRACT

Qualitive data from interviews with individuals and observations of group meetings were gathered from 16 school districts that participated in a study of school district uses of evaluation and test data. The findings are organized around a series of discrete topics presented in a set of independent papers. The title essay, "Working Knowledge," offers a brief note on how illustrative quotes were selected for presentation, elaborates on the concept of working knowledge, and discusses the interaction between working knowledge and evidence. "Evidence and Thought" reviews 7 examples of collective conceptual uses of evidence selected from a pool of 43 episodes. "Evidence and Decision" reviews the details of 14 decisions that involved evidence that at least some participants claimed was instructive to the decision. "Evidence and Management" provides an evalation of management strategies loosely grouped at four points along a continuum according to the amount of emphasis they placed on tests. "The Role of the In-House Evaluator" describes the roles that evaluation units had adopted and how these roles fit into their organizations. Each paper is followed by references. The appendixes contain sampling and data collection procedures; profiles of the 16 school districts; and an analysis of claims about the use of evidence. (MLF)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

The Huron Institute

EA

ED248605

WORKING KNOWLEDGE

And Other Essays

Mary M. Kennedy

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION
CENTER (ERIC)

☒ This document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

- Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

EA 017 143

THE HURON INSTITUTE 123 MOUNT AUBURN STREET, CAMBRIDGE, MASSACHUSETTS 02138

WORKING KNOWLEDGE

And Other Essays

Mary M. Kennedy

The Huron Institute
123 Mt. Auburn St.
Cambridge, MA. 02178
(617) 491-5450

September, 1982

PREFACE

This is the final report of a research project commissioned by the National Institute of Education (NIE) in 1979. When NIE solicited this project, it was concerned about the extent to which school districts were able to tie evaluation and test data to managerial, instructional, or programmatic improvement. Though federal policy interests have changed considerably since that time, the issues raised by the NIE are still pertinent. Not only federal policy, but many state and local policies as well require the production and use of evaluation or test data. These policies are based on the assumption that such data can and should be used to improve educational practices. Findings from this study are therefore relevant to managers at all levels of educational governance -- federal, state and local -- as well as to educational researchers and evaluators who are interested in promoting the use of such evidence for educational improvement.

Recognizing, however, that many of these audiences do not have the time to read lengthy reports, I have chosen reporting strategies that would increase the accessibility of my findings to busy readers.

My first strategy is to organize the findings around a series of discrete topics rather than arranging them into a single overall report. The papers presented in this volume are independent of one another and can be read in any order. This strategy was chosen in order to better serve readers who are concerned about only one or two topics.

My second strategy is simply to be brief. Brevity is especially difficult to achieve when reporting qualitative data, for the data themselves are voluminous and they cannot be summarized as automatically as quantitative data can. I therefore devoted considerable energy to creating analytic techniques that would enable me to collapse and summarize these data, and as a result I have been able to reduce over 3,000 pages of data to a report of roughly 200 pages. And I think I have done so without sacrificing either depth or breadth of coverage.

Also in the name of brevity I have avoided lengthy reviews of the literature and have restricted the number of references I make to the literature. This decision contains the lengths of the papers and enables the space that is used to be heavily saturated with new data rather than with old. Though the papers themselves do not contain lengthy discussions of the literature, they are nonetheless influenced by the literature, and I take this opportunity to acknowledge those authors who have been most influential: David K. Cohen, Charles E. Lindblom, James G. March, Martin S. Rein, and Carol H. Weiss. None of these authors is, of course, responsible for the papers presented here or for the way their ideas have been interpreted here.

CONTENTS

WORKING KNOWLEDGE 1

A Brief Note on Choosing Illustrative Quotes 5
Characteristics of Working Knowledge 8
The Interaction Between Evidence and Working Knowledge 13
Conclusion 25

EVIDENCE AND THOUGHT 29

Diverse Responses to Evidence 36
Attempts at Unifying Responses to the Evidence 45
Conclusion 53

EVIDENCE AND DECISION 59

The Decision-Making Context 69
The Decisions 76
Summary and Conclusions 96

EVIDENCE AND MANAGEMENT 105

Consultation 110
Instigation 113
Regulation 115
Evaluation 120
Conclusion 123

THE ROLE OF THE IN-HOUSE EVALUATOR 127

Organizational Influences on the Evaluator's Role 130
Problems of Adaptation. 137
Effects of Successful Adaptation 142
Summary and Conclusions 161

APPENDIX A 165
Sampling and Data Collection Procecures

APPENDIX B 179
Profiles of the 16 School Districts involved in the Final Analysis

APPENDIX C 199
The Analysis of Claims about the Use of Evidence

WORKING KNOWLEDGE

Much of the literature on how social science evidence is used tends to be concerned with its use in formal decision-making situations. Public policy makers or administrators are envisioned as acting something like a jury, which has a clear body of evidence it must use, clear rules for how that evidence should be weighed, and a clear time and place in which the decision must be made. The analogy is not perfect, of course, because policy issues ebb and flow, change and circle back again. Within the overall process, there are particular decision points, such as times when all participants are expected to vote, but in between these times are hundreds of occasions whose boundaries are not clear, but which may nevertheless influence policies or practices. These are the hallway conversations, the hearings, the committee meetings and so forth, when participants must spontaneously draw on whatever knowledge is in their heads to respond to whatever ideas have been put forward.

Working knowledge is the organized body of knowledge that administrators and policy makers use spontaneously and routinely in the context of their work. It includes the entire array of beliefs, assumptions, interests, and experiences that influence the behavior of individuals at work. It also includes social science knowledge. The term _working_, as used here, has two meanings. First, it means that this is a special domain of knowledge that is relevant to one's job. Second, it means that the knowledge itself is tentative, subject to change as the worker encounters new situations or new evidence. Although administrators and policy makers may prepare for particular decisive events by studying relevant social science evidence, they must still depend on their working knowledge for the majority of situations they encounter. Working knowledge often has a greater cumulative influence on policies and practices than does the evidence that is specifically brought to formal decision points.

Despite the convenience and broad applicability of working knowledge, there are reasons to distrust the quality of judgments and decisions that are based on it. Cognitive psychologists have documented a wide range of weaknesses and flaws in unaided human thought processes (Faust, 1982; Kahneman, Slovik and Tversky, 1982; Kaplan and Schwartz, 1975; Meehl, 1971; and Sadler, 1981), and have suggested that clinical insight is not nearly as powerful as those who use it would like to think. Findings such as these are among the reasons why some social scientists feel that social science should play a greater role in the decision making process. Hammond (1978), for instance, defines six "modes of inquiry," which differ primarily in the extent to which they rely on scientific evidence as opposed to private judgments. The sixth mode is most analogous to ad hoc uses of working knowledge. Hammond describes this form of reasoning as

> . . . the kind of thought most of us engage in most of the time. It involves an uncertain data base, no manipulation of variables, no statistical controls, and inconsistent logical rules never made explicit. . . . [It is] particularly vulnerable to the effects of numerous psychological factors and therefore it is methodologically very weak. Moreover, . . . no one (not even the person making the judgment) can be sure of what the judgment process is. . . . In short, *[it] is not only the weakest means for solving problems, it is the most dangerous one*. [1978, p. 18, emphasis added]

What is not clear, however, is whether or to what extent the availability of social science evidence would improve these ad-hoc judgments. To suggest that it would is to make two important assumptions: first, that social science offers a superior form of reasoning, as well as a superior form of knowledge; and second, that the policy maker or administrator has a choice about what knowledge will be used. These assumptions do not apply to the thousands of daily situations when the policy maker has access to nothing but working knowledge. For these situations, social science can only be used if it has become a part of working knowledge, so that the findings of relevant research are readily available to the user.

Little attention has been paid to the relationship between working knowledge and social science evidence -- to how evidence is incorporated into working knowledge and to how these two structures of knowledge influence one another. There are two possible directions of influence. On the one side, Weiss (1977, 1980) has shown that evidence feeds into working knowledge, expands it, and can have a major role in changing it. This is the direction of influence that social scientists wish to encourage. But on the other side of the relationship, working knowledge is used to interpret new evidence and to judge the validity and applicability of each new source of evidence encountered (Lindblom and Cohen, 1979). This side of the relationship is also important, for social science comes in both good and bad forms, and the range of evidence available must be sifted to determine what is valid and what is relevant. However, given the limited capacity of the human brain for synthesizing complex bodies of data, serious errors could occur in the process of integrating these two bodies of knowledge. This paper is designed to shed light on these issues by describing the way in which evidence becomes encorporated into the working knowledge of public school administrators and teachers.

The data on which this paper is based came from 16 school districts which participated in a study of school district uses of evaluation and test data. The districts were quite diverse. They ranged from poor to wealthy, served from 4,000 to 240,000 students, served communities in all regions of the country, and their student bodies ranged from mostly white to mostly black to mostly Hispanic. The data gathered from these districts were entirely qualitative, coming from interviews with individuals or from observations of group meetings. Although the observations were limited to those meetings which happened to occur at the time of our field work, the interviews were scheduled to include members of the policy-making community (usually superintendents, assistant superintendents or school board members), the program development community (usually program directors, curriculum coordinators and supervisors), school buildings (usually principals) and classrooms (teachers).

The intent behind both observations and interviews was to expose the relationship between evidence and the working knowledge these participants had about substantive issues within their districts. Observers described everything that transpired during meetings, including any references to evidence, so that their notes from these meetings could indicate the substantive context in which the evidence was drawn upon. Interviewers discussed issues of current interest to interviewees, rather than the use of evidence per se, but they did so with an eye toward documenting how and where different kinds of evidence fit into the interviewee's train of thought, if it did at all.

Analyses of the notes from these observations and interviews indicate that there are three analytically distinct, though in practice interdependent, processes involved in the use of evidence. The first is that of seeking out new evidence and attending to it; the second is that of incorporating it into existing working knowledge, and the third is that of applying it to working situations as they arise. The processes are dynamically interdependent in that all of them contribute to the ongoing evolution of thought and action. Since this paper addresses only the relationship between evidence and working knowledge, it presents findings only about the first two of these three processes.

The paper has three main sections. The first offers a brief note on how illustrative quotes were selected for presentation. The second elaborates on the concept of working knowledge and the third discusses the interaction between working knowledge and evidence.

A BRIEF NOTE ON CHOOSING ILLUSTRATIVE QUOTES

The argument presented here relies heavily on an analysis of verbal material gathered either from interviews or from observations of meetings. All references to evidence that occurred in the notes were taken from the notes and sorted according to their content. Table 1 presents the results of this analysis. The upper box includes a total of 728 citations, and these form the basis of this paper.

One of the problems inherent in reporting the results of such an analysis is that the reporter can not present averages, but instead must present illustrative material. Yet readers may not be sure how typical these examples really are. Here, then, are the rules I followed for selecting illustrative citations.

TABLE 1

Summary of All Comments Pertaining to the Use of Formal Information

Contents of Citation	Context of the Comment				Total Number of Citations
	Policy Issues	Program Issues	School Issues	Classroom Issues	
SEEKING INFORMATION					
Process of looking	25	15	36	24	100
Rationale for looking	92	105	104	94	395
PERSONAL LEARNING					
Descriptive Knowledge	31	21	23	7	82
Inferences or Conclusions	50	37	40	24	151
Total	198	178	203	149	728
USE IN OTHER WORKING SITUATIONS	165	153	139	124	581
TOTAL OF ALL CITATIONS	363	331	342	273	1309

1. I have restricted myself to those examples that are relatively self-explanatory, avoiding those that are highly ideosyncratic or particularistic, and consequently require more contextual description in order to be understood.
2. I have avoided illustrations that are too short or lacking in detail, such as, for instance, comments to the effect that test data are "helpful" or that a study was "informative," as well as those that are too long. Ruling out overly long illustrations effectively means ruling out observation material, for although much of the observations vividly illustrate the points made here, they tend to be more complicated examples and therefore are difficult to quickly summarize.
3. I have avoided examples in which interviewers paraphrased their interviewees, rather than directly quoting them. This decision meant that some districts could not be called upon as often as others, since field workers varied in their inclination to directly quote their respondents. I have, however, checked the relative frequency with which different kinds of examples appear across districts, and found no evidence of variation among districts on the points made in this paper. The preference for direct quotes is entirely an aesthetic one.
4. Within a given set of examples, I attempted to illustrate variation rather than typicality, and I sought variation on these dimensions and in this order:
    - First, I tried to vary the substance of the comment
    - Second I tried to vary the districts from which the examples came, unless the point of the presentation is to illustrate

within-district variation

- Third, I tried to vary the titles of the people quoted.

5. Across the sets of examples, I also tried to vary districts and positions of interviewees.

The first three of these rules are more valuable to the reader than to the analyst, for they enable a more succinct and lively presentation of the findings. The latter two rules are more valuable to the analyst, for they force me to check my impressions of trends against the full set of data, rather than relying on those districts I personally visited and consequently know best. The result of these rules is that the illustrations do not in fact represent the full set of data in the way that a random selection would. But they do not misrepresent it with respect to the points made in this paper.

CHARACTERISTICS OF WORKING KNOWLEDGE

Human beings are apparently capable of collecting and organizing an amazing variety of information into global, if somewhat vague, patterns. How this is done has been the topic of research in psychology, sociology, and organizational theory. And the products -- the organized systems of knowledge that result -- have been called gestalts (Kohler, 1970), cell assemblies (Hebb, 1949), schemata (Piaget, 1971), problem spaces (Newell and Simon, 1972), theories of action (Argyris and Schon, 1978), conceptions of social reality (Caplan, 1975), and Weltanschauung (Weiss, 1980). Broadly speaking, these several terms refer to the same phenomenon of actively organizing knowledge, but each has its own special meaning that limits its applicability to a particular subset of the vast range of knowledge people normally have. Gestalts, for instance, are perceptual configurations that are formed almost instantaneously upon perceiving a situation. The whole gestalt cannot be derived from merely a listing of the parts, for it is the relationship among the parts that is significant, rather than the parts themselves. Schemata, on the other hand, are representations of dynamic properties of physical objects, particularly as they counterbalance one another. They are not acquired as quickly as gestalts -- we have to manipulate a situation in order to develop a schema of it. For instance, our knowledge of teeter-totters constitutes a schema. We know that both ends cannot be up simultaneously, presumably because we have teetered on them or watched others teeter on them, and these active experiences led us to construct schemas of teeter-totters that represent them in a particular dynamic way. Problem spaces are different still. They are definitions of particular problems. A problem space includes the variables involved in the problem, the heuristics and algorythms that may be used to solve it, and some estimates of, or assumptions about, what the solution might look like. People create unique problem spaces for each problem they face. Theories of action encompass yet another aspect of knowledge. They are developed from social experiences, and include such things as estimates of other people's points of view and of how others might respond to one's own behavior.

Working knowledge does not quite fit any of these definitions, though it contains elements of them all. Working knowledge is more subject to change than are gestalts and schemas, it contributes to the full range of problems people encounter at work, rather than being limited to a particular problem, and it includes more than just what is learned from experiences at work. For participants in this study, working knowledge included assumptions about how children learn and develop, pedagogical theories and educational philosophies, legal and economic knowledge, knowledge about how educational services in their districts were organized and delivered, about how well certain programs or certain colleagues were performing, and the interests and predilections of their colleagues, as well as value judgments about all these things, goals pertaining to them and consequent interests in them.

If working knowledge were to be broken down into its constituent parts, four distinct components could be identified. Two of these, formal evidence and experiences, constitute the empirical parts of working knowledge. The other two parts are the individual's interests, or goals; and his beliefs, which include myths and legends as well as value judgments. But to say that such components can be identified is not to say that any particular statement made by an interviewee could be labeled as belonging to one category or another. What appears to be myth could in fact be based on evidence, and what appears to be based on evidence could be myth. Some authors have suggested that individual interests have a dominant role in the overall system of knowledge. Holzner and Fisher (1979), for example, argue that knowledge is organized according to its intended uses, and Lindblom and Cohen (1979) argue that it will necessarily be used to serve one's own interests.

In the minds of users, the components of working knowledge are in fact indistinguishable. They are blended together to form an integrated and organized body of knowledge. The result can be seen in the following example. One school board member participating in this study was interviewed at length about her views regarding whether or not her district should convert junior high schools to middle schools. During the course of the interview, she listed a number of reasons why she opposed this idea. The variables she considered included the costs of the conversion and the probable effects it would have on student achievement, student social development, and student exposure to drugs. She viewed the issue from a number of angles and her conclusions had been formed from a variety of elements of her working knowledge. Throughout the interview, she brought in a number of facts to support her arguments, and the interviewer repeatedly asked her where she had learned these things. Her sources included a newspaper article describing the results of a comparative study done in another school district, a survey of parent attitudes conducted locally, her own observations, things teachers and parents had told her, local budget documents, "common sense" and so on. Toward the end of the interview, on being asked once again how she knew something, she said the individual facts were not as relevant as the collection of them was:

> You see, it's piecing it all together. For instance, we get kids pulled out for drug abuse. And the ages seem to be getting younger and younger. You have two things occurring at the same time. One, kids are exposed to more things, especially at junior highs. . . . [Second], I can break that down into feeder schools -- kids from an elementary building with a strong anti-substance abuse program are less likely to get into drugs. This helped quite a lot. But I can still pick up my drug instance report and see sixth graders. [District 115,[1] school board member]

In addition to being characterized by its component parts, its organization and its continual evolution, working knowledge can also be characterized by its individuality. Even members of the same district can encounter different situations and different evidence, thus developing different bodies of working knowledge and "improving" their programs in different, even opposite, ways. For instance, in district 220, two program directors came to opposite conclusions about how to supply resource teachers to help regular teachers implement their programs. One said, "I'll get eight people to travel so they won't get involved in local stuff. With my current staffing, resource teachers visit the schools too much and they get involved in local crap. [How do you know that?] I know the people and I know the situation, and I get weekly reports from everybody. There's a comment section in these reports and I can tell what business they're getting into by the nature of their comments and concerns" [District 220, director of reading curriculum]. The other program director said he got the idea for his program from a model he had seen "out West somewhere," and that the people there told him it didn't work unless the resource teachers were accepted by the building staff. He opted for a residential expert in each building, saying, "They're part of the local gang; they're not outsiders who are sent in" [District 220, director of math curriculum].

This then is working knowledge. It is continually accumulating and evolving; it consists not only of evidence but of experiences, interests and beliefs as well; it is organized; and its contents and organization differ from one individual to another.

THE INTERACTION BETWEEN EVIDENCE AND WORKING KNOWLEDGE

The foregoing account of working knowledge suggests not only that it can change over time but also that evidence can be a part of it. It does not, however, indicate how evidence and working knowledge accommodate one another. This section describes two interrelated processes that are relevant to that interaction. First is the process by which users seek out new evidence and second is the process by which they incorporate new evidence into their working knowledge.

Seeking Out New Evidence

Participants' descriptions of the process of seeking out new evidence suggest that the process could be characterized by three adjectives: active, continual, and unsystematic. The process is active in that participants usually did more than merely glance at reports that came their way. Though they often perused them very quickly, they did pay attention to them, and they looked for information that they thought might contribute to their understanding of their environment. It was continual in that new evidence was continually becoming available, and it was unsystematic in that participants tended to look indiscriminately at everything that came their way, and in that they could not describe exactly what it was they were looking for. Here are some descriptions of the process.

- [showing us histograms he has put together himself] I put these together every year when the test data come out, and I use them to talk with teachers about the strengths and weaknesses for each area the test measures. [District 50, elementary principal]

- When I look at test scores, I look for patterns. . . . I look for trends and red flags. [District 4, superintendent]

- I'm interested in looking at the difference between math and reading [in the computer-assisted program]. I don't have any ideas about whether the program works better in one area or another, but it would be interesting to see. [District 25, director of bilingual education]

- What happens is you look at the summary and something will catch your eye, so I will asterisk that and ask for a review of that. [District 220, director of Title I programs]

Though these participants were aware that their search was active and continual, they would probably not have called it unsystematic. They knew what they were looking for -- strengths and weaknesses, trends and red flags. Something that would catch their eyes. Yet even when they gave their reasons for studying new evidence, the reasons were vague.

- I find evaluation reports useful in stimulating me to think about the curriculum in new ways. [District 50, program director]

- It seems to me that this [problem of differences in test scores between black and white students] is an area we have to start looking at. If we don't publish these data, we won't even have a definition of the problem. [District 115, assistant superintendent]

- I always have an eye out for new ideas for bringing money into the district for the arts, because "basic skills" is threatening our program. [District 35, director of arts]
- [Regarding intelligence quotient scores] If a child is doing poorly I can see if there is a problem. It gives you a place to start. [District 27, teacher]

For these participants the purpose of the search was not to find answers to particular questions, nor to solve pressing problems. In that sense, the search was not systematic. But it was nevertheless controlled by their interests and by that large and vague body of working knowledge that served those interests.

These comments suggest two things regarding the relationship between working knowledge and evidence. First, participants appear to be aware that their knowledge is incomplete or inadequate, and that they need evidence. Second, participants seem to be genuinely open to the knowledge that can be gained from formal evidence. They are not only willing to, but also want to, use evidence to expand or refine their working knowledge. And they expect evidence to influence their working knowledge. But in order for evidence to have such an effect, it must become a working part of their knowledge. Participants must do more than study it; they must incorporate it into their knowledge.

Incorporating New Evidence

Incorporation is the process of making evidence a part of working knowledge. Though participants in this study often claimed to have learned from the evidence they reviewed, these claims were not automatically taken as valid. For a comment to be considered a valid illustration that something was learned, it had to meet two criteria. First, it had to state what the specific knowledge was, rather than stating that evidence was, say, helpful, interesting, informative, etc. Second, the comment had to be generated spontaneously during the course of a discussion about a substantive issue, rather than given in response to a specific question about the use of evidence. Using these two criteria, I found 233 examples in which formal evidence had been incorporated into working knowledge. These examples indicated that incorporation could occur in three different ways. First, the evidence could be incorporated in its original form. This often occurred when the evidence consisted of simple descriptive statistics. Second, it could be interpreted. When this occurred, the interpretation as well as the evidence itself became a part of working knowledge. Third, inferences could be drawn by bridging the evidence to already available working knowledge. The inference actually derives from the bridge, rather than from the evidence alone or from working knowledge alone. In these cases, the bridging inference was also incorporated into the body of working knowledge. All three of these constitute cases in which evidence has influenced working knowledge, but they are also cases in which working knowledge has influenced the evidence.

Here are some examples of comments which indicated that evidence was retained in its original form.

- Between 60 and 65 percent of the citizens in this community do not have school-aged children. [District 35, superintendent]
- This school has changed from 80% minority to 50% minority in just the past three years. [District 50, elementary principal]

- o Special education enrollment is not decreasing even though regular education enrollment is. [District 4, director of special education]
- o We have the brightest fifth grade we've ever had. [District 7, assistant superintendent]

The volume of evidence from which these participants could learn such facts is enormous. Every district in this study had multiple annual testing programs and had annual data on enrollment and attendance, and many had annual data on vandalism and drug abuses as well. Every district had annual evaluations of its state- or federally-funded programs since these evaluations were mandatory. Many had findings from surveys of high school graduates or of the community, surveys required by accrediting agencies. And those districts with their own evaluation offices had other studies and evaluations as well. Particular facts are not randomly recalled from this mass of information; they are recalled because they are meaningful to the speaker. The superintendent who notices that the majority of the citizenry does not have school-aged children is also aware of the implications of that fact for the school's enrollment and for whether future tax millage increases will be approved by the voters. And the school principal who can describe the changes in his building's enrollment pattern can see the effects of this change every day when he walks his hallways. Without a method of screening the mass of available data, the entire array might be incomprehensible. Working knowledge provides a means of identifying those facts that are relevant, and once identified, those facts then become part of working knowledge.

The above illustrations suggest not only that the participants incorporated those descriptive statistics that were relevant to them, but also that the facts were accurately retained, and there is the possibility that the latter point is not true. Since I could not find all of the original data to which participants referred, I could not judge the accuracy of many of these remarks. However, I did encounter some evidence of inaccuracies in cases where the same evidence was referred to by multiple interviewees in the same district. For instance, district 7 had conducted two surveys, one fifteen years prior to this study and the other, one year before this study. Both were done to fulfill accreditation requirements for the district's only high school, and both were interpreted by members of the district as indicating a need for more vocational education. Here are five references to these data:

- [Fourteen years ago,] only 40 percent of the high school students were going to college. And of that group, only 20 percent were graduating from college. We [therefore] felt we needed to teach students salable skills. [Superintendent]

- Now only 40 percent of the middle school students plan [to go to] college. So we'll need to make some curriculum changes and offer more vocational education. [Associate superintendent]

- We realized in the early sixties that only 55 percent of the students were going on to college. We sold the board on the idea of vocational education. [Senior High Principal]

- It used to be that parents of students here, 40 percent of them had been to college, but now only 17 percent of our students' parents have been to college. [Vocational education coordinator]

These participants all knew why the evidence had been relevant: the statistics indicated a need to expand the high school's vocational education program. But participants were less sure of what the facts actually were, perhaps because the particular statistics were less important to their working knowledge than the implications were.

These examples from district 7 introduce the second way in which evidence can be incorporated into working knowledge: facts are not merely recognized as relevant, but are made relevant by being interpreted. Here are some examples in which evidence has been interpreted.

- [In reference to the migration of families in and out of the district] We still lose about 25 percent of our enrollment before the end of the school year, but we used to lose 90 percent. The migrant parents' views about education are beginning to change. They're starting to realize that they should wait until the end of the school year before leaving. [District 25, principal]
- The mobility of students in this district is evidence of the need for a curriculum that is uniform across the whole district. [District 83, curriculum director]
- I saw math scores go up after the change. Reading scores are down slightly but are solid. [District 50, Title I director]

Keeping in mind that none of the figures quoted to interviewers may be accurate, examination of these interpretations sheds some light on the relationship between evidence and working knowledge, for the interpretations rely on other aspects of already available working knowledge. Individual interests influence many interpretations. The curriculum director quoted above who interpreted student mobility as indicating a need for a uniform curriculum would gain a great deal of influence if his district mandated a uniform curriculum. Other examples of personal interests influencing interpretations include a special education program director who had just completed her master's degree in the education of the deaf and interpreted the test scores in her district as indicating a need for a new program for deaf children [District 240]; a principal in a school district with only five elementary schools describing her school's test scores as "third from the top" in the district, rather than third from the bottom [District 4]; and the director of a bilingual education program who had just received a very negative evaluation of her program and interpreted the evidence not as indicating that the program's practices were inadequate, but rather that,

> we were expecting too much from our students and I think if we can be more realistic we can accomplish what we set out to do. [District 57]

Beliefs also play a part in interpreting evidence. A school board member uses value judgments when he says, "Eight hundred is too many to close a school" [District 115], and a principal draws on her values when she says she is dismayed by the findings from a study that indicated that a disproportionate number of minority students (boys) receive corporal punishment or are suspended [District 50]. There were also some cases in which participants appeared to use myths to interpret their evidence.

- [In reference to a lack of test score growth among the district's sixth graders] It's a national phenomenon that when kids reach sixth grade there is a plateau. [District 7, superintendent talking with his cabinet]

- [In reference to a change in first grade scores from the thirty-second percentile to the seventy-second] I brought teachers in to discuss this. I know test scores are up all over the country. [District 240, principal]

Sometimes the particular aspects of working knowledge that are used to interpret evidence cannot be identified, even though the interpretations themselves can be. For instance, two principals in the same school district received test data indicating that children in the first and second grades were scoring relatively high, while children in the latter grades were scoring relatively low. Here is how each of them interpreted these data.

- Our kids don't do as well in later grades as in earlier grades. I think it's a function of the test and to some extent the curriculum. I think the test is more biased for our population in the later grades. [District 220, principal; emphasis added]
- When I get the test results I do nothing with them. . . . I know the results are inflated for grades one and two. So I don't put much reliance on the first grade scores because they just don't jibe with what we're doing. [District 220, principal; emphasis added]

Interpretations, then, rely on elements of already available working knowledge. Working knowledge enables users to define the relevance of the evidence to their working situations, thereby rendering the evidence meaningful. But meaningful evidence is not necessarily accurately retained, even though its meaning is retained. The administrators in district 7 could all recall their interpretations of the evidence, but when called upon to justify their conclusions, they appeared to reconstruct factual data that were consistent with their original interpretations.

The third set of examples includes those cases in which bridging inferences are drawn. These inferences connect evidence to other elements of working knowledge -- interests, experiences or beliefs. For instance, a teacher uses an inference to build a bridge between his experiences and his classroom's performance on test scores:

> Last year I was the lowest in the sixth grade and this year I'm the highest, and I've done nothing different. . . . Tests are simply not reliable. [District 220, teacher]

And a board member bridges two kinds of evidence with an inference.

> The achievement test results indicate a great diversity in our student body. Twenty-two percent of them are Oriental. [District 4, school board president]

Bridging inferences not only enable users to view the evidence as part of a larger picture of what is occurring, but also to view it as part of a larger argument for what should be occurring. A school board member in district 220 discussed the importance of parent involvement in education and bolstered his argument with the following facts: (a) there is a known relationship between student attitudes and parent attitudes, (b) low test scores tend to be in schools with low attendance, and (c) one school with severe attendance problems served students for whom only 12 percent of the parents had ever completed high school.

He could not recall the sources for these facts, even for the specific 12 percent figures, and said, "I don't know how I know that; it's from something I read somewhere" (District 220, board member). But even if these "facts" are all true, they do not automatically add up to a conclusion that parent involvement is needed. He has bridged these relationships together to create a model that runs something like this: parent attitudes affect student attitudes, student attitudes affect student attendance, and student attendance affects student's grades. Parent involvement in the educational process can improve parent attitudes toward education and can thereby improve their children's attitudes, attendance, and ultimately their achievement as well.

Most bridging inferences are implicit rather than explicit, and must themselves be inferred by interviewers. For instance, one superintendent had been impressed by a study of voter behavior which indicated that senior citizens voted out of proportion to all other groups of voters. He said,

> This study showed me two things. It showed me the weak links [in my campaign to get voter approval for a proposed tax millage increase for the schools] and it made me study the location of our young families and senior citizens. . . . You've got to be aggressive! [District 7, superintendent]

Clearly the study did not tell this superintendent to be aggressive, nor did it say anything else about what he should or should not do. The study was not even conducted in his community. But he did seek it out and read it with the intention of bridging it to his knowledge of his own situation. The inferences he drew were built by bringing a number of elements in his working knowledge, no doubt including the following: (a) *experiences* or *beliefs* to the effect that senior citizens had less money to contribute to taxes than working citizens did and that senior citizens did not have children in the schools; (b) an *inference* that senior citizens would be less likely to approve of a tax millage increase for the schools than employed parents of school-aged children would be; (c) *legal knowledge* that the majority of people who actually vote must approve of a bill in order for it to pass; (d) *evidence* that senior citizens usually comprise the majority of voters in elections; (e) *interest* in having the millage bill pass; (f) *experiences* of the campaign he had waged so far, which was targeted primarily toward parents of children enrolled in his schools; and (g) the *inference* that even if he convinced the majority of parents to approve of the bill, these parents might not constitute the majority of voters. The three inferences he volunteered during his conversation were the bridges that connected all of these elements of his working knowledge together: (h) his campaign had weak links in that it did not include senior citizens; (i) he needed to learn more about the voting residences of senior citizens and working parents in his community; and (j) he needed to be more aggressive in order to succeed in his campaign.

When participants describe their processes of seeking out evidence and give their reasons for looking at evidence, they indicate both an awareness that their knowledge is tentative and a willingness to use new evidence to improve it. Their comments imply that the evidence at least has the potential to greatly influence their working knowledge.

But when they relate what they have actually learned from the evidence, they tell a slightly different story. It is not clear that even the most rudimentary descriptive statistics influenced working knowledge without first being influenced by working knowledge. And the most substantial alterations in working knowledge came from interpretations of evidence and from bridging inferences between evidence and other elements of working knowledge, not from the evidence per se.

CONCLUSION

Advocates for a stronger role for social science evidence in public policy and decision-making often assume that such evidence can help public administrators and policy makers improve both their knowledge and their reasoning processes. But before such evidence can contribute, it must first be comprehended by its users. Working knowledge is the link between evidence and its application. Once evidence is discovered and incorporated into working knowledge, it becomes part of the user's conceptual baggage and is taken to all working situations and used for any of the variety of things working educators do. But evidence is not merely attached to working knowledge like barnacles are attached to clams, riding unchanged from one situation to another. Rather, it is acted upon by working knowledge. It is sorted, sifted, and interpreted, and its original source and character are often lost. Even descriptive facts are changed, as all evidence is translated into implications, interpretations and bridging inferences that enable users to organize their store of working knowledge. It is these interpretations and bridging inferences that become part of working knowledge, and it is these, rather than the evidence per se, that are carried to working situations and used.

NOTES

[1] District code numbers indicate the approximate number of students served in thousands. District 4 serves about 4,000 students and district 240 around 240,000 students. Code numbers deviate randomly from actual enrollments by ±15 percent.

REFERENCES

Argyris, C. and Schon, D. A. Organizational learning: A theory of action perspective. Reading, MA: Addison-Wesley, 1978.

Caplan, N. A., Morrison, A. and Stambaugh, R. J. The use of social science knowledge in policy decisions at the national level: A report to respondents. Ann Arbor, MI: University of Michigan,

Faust, D. A needed component in prescriptions for science: Empirical knowledge of human cognitive limitations. Knowledge: Creation, Diffusion, Utilization, 1982, 3, 555-570.

Hammond, K. R. Toward increasing competence of thought in public policy formation. In K. R. Hammond (Ed.), Judgment and decision in public policy formation. Boulder, CO: Westview Press, 1978.

Hebb, D. O. Organization of behavior. New York: John Wiley and Sons, 1949.

Holzner, B., and E. Fisher. Knowledge in use: Considerations in the sociology of knowledge application. Knowledge: Creation, Diffusion, Utilization, 1979, 1, 219-244.

Kahneman, D., P. Slovik, and A. Tversky (Eds.) Judgment under uncertainty: Heuristics and biases. New York: Cambridge University Press, 1982.

Kohler, W. Gestalt Psychology. New York: Liverright, 1970.

Lindblom, C. E. and Cohen, D. K. Usable knowledge: Social science and social problem solving. New Haven, CT: Yale University Press, 1979.

Meehl, P. E. Law and the fireside inductions: Some reflections of a clinical psychologist. Journal of Social Issues, 1971, 27(4), 65-100.

Newell, A. and Simon, H. A. Human problem solving. Englewood Cliffs, N. J.: Prentice-Hall, 1972.

Piaget, J. Biology and knowledge. Chicago: The University of Chicago Press, 1971.

Sadler, D. R. Intuitive data processing as a potential source of bias in naturalistic evaluations. Educational Evaluation and Policy Analysis, 1981, 3(4), 25-31.

Weiss, C. H. Research for policy's sake: The enlightenment function of social research. Policy Analysis, 1977, 3(4), 531-545.

Weiss, C. H. Knowledge creep and decision accretion. Knowledge: Creation, Diffusion, Utilization, 1980, 1, 381-404.

EVIDENCE AND THOUGHT

Though social science can be and is used in a variety of ways, two kinds of use have received more research attention than others -- so much so, in fact, that models have been gradually developed to portray the idealized versions of each kind of use. One of these is the _instrumental_ model. Under this model, "use" consists of making a decision, and social science evidence is assumed to be instructive to that decision. To the extent that a choice, for instance, is based on evidence of the costs and benefits of alternative courses of action, the choice is considered to be sound, or rational; it is a better decision because of its reliance on the evidence. Though there are variations on this theme, the model usually assumes that once the evidence is available, the decision is relatively straightforward. All that is needed for the decision is the right evidence. The other model of use could be called the _conceptual model_. Under this model, evidence is not specifically instructive, but it is nevertheless relevant, and "use" consists of thinking about the evidence. Although instrumental use is generally construed to be one particular thing, conceptual use can be any of a variety of things. It can, for instance, consist of discovering that one is laboring under false assumptions and altering one's perceptions accordingly, of developing a broader, deeper, or more sophisticated understanding of issues, of confirming suspicions, providing new insights, challenging assumptions, or in other ways changing one's ideas about one's policies or practices or about the social problems one's policies or practices are designed to affect.

The most important difference between the conceptual model and the instrumental model is this: Whereas the central feature of the instrumental model of use is the decision, the central feature of the conceptual model is the human information processer, and that fact leads to several other important distinctions between this model and the instrumental model.

For instance, whereas the decision is assumed to consist of nothing more than a set of options waiting for the evidence to sort them out, the human information processer is assumed to already have a considerable body of knowledge and ideas before receiving the evidence. It approaches the evidence with a well-developed, internalized model of the substantive issues at hand. If, for instance, the policy maker works primarily in education, he or she approaches new evidence with already-developed ideas about child development and the learning processes; about how teachers operate, what motivates them and how to tell a good one from a bad one; about the different kinds of effects that different kinds of policies or programs are likely to have and why; and about how well the administrators of existing educational programs are doing their jobs. These ideas are arranged in an organized coherent body of working knowledge within which individual policy makers and administrators operate. The presence of this working knowledge is essential to the conceptual model. Since evidence is no longer instructive to a decision, as it was in the instrumental model, it must have a bearing on something else. In the conceptual model, it influences working knowledge.

Second, whereas the instrumental model posits the decision as a passive recipient of evidence, something which responds almost automatically to the instructions contained in the evidence, the conceptual model posits its information processer as active. Rather than responding automatically to the evidence, the processer interprets its meaning, decides its relevance, and hence determines how the evidence will influence its own working knowledge. The information processer actively renders the evidence meaningful, and in that sense it influences the evidence before it is influenced by the evidence. The notion that evidence is "used" conceptually, then, means that it is acted upon conceptually.

The practical implications of the conceptual model are also less understood. Authors such as Deising (1962) and Allison (1971), for instance, have fully elaborated the nuances of rational decision making, from which the instrumental model of use derives. But although several authors have suggested the idea of conceptual use (e.g., Weiss, 1977) or the idea of a human information processer (e.g., Caplan, 1977), none has specified the model in much detail. Thus, although the model is sufficiently developed to at least describe a plausible process by which evidence could be used, some aspects of the model have not been fully explored. One of these has to do with whether and how evidence might influence practice when it is used conceptually, a second relates to the problem of whether and how groups of people develop agreed-upon interpretations of the evidence, and a third relates to the fidelity of interpretations relative to the evidence itself.

Leviton and Hughes (1981) and Rich (1977) have suggested that conceptual use be defined in part by its lack of visible effects on practice. This criterion clearly distinguishes conceptual use from instrumental use, and it may be a useful one for research purposes, since whatever effects conceptual use of evidence has on practice may be far less straightforward than the effects instrumental use has on practice, and consequently be harder for a researcher to discern. But if no changes in practice were to occur at all, if only thought were to change, then it would be difficult to argue that the evidence was really beneficial. Ultimately, even if time has passed, even if the changes are only circuitously related to the evidence, even if the contributions of the evidence are delayed, obscure, and indirect, they should still be there if the evidence is to be said to have had practical value. And effects of conceptual use on practice have been observed to occur in a number of different ways. Evidence can tilt a policy-maker's point of view so that a series of small decisions and actions are performed differently than they otherwise would have been (Alkin, Daillak and White, 1978); it can accumulate with other evidence and other sources of knowledge and ideas to influence future decisions and actions (Weiss, 1980); or it can stimulate policy makers to perceive the issues differently in the future or lead them to define their options differently than they would have if they had not refined their understanding of the issues (Cohen and Weiss, 1979). If researchers accept changes in thought alone as sufficient criteria of conceptual use, it is only because such changes have at least created the possibility for future changes in practice, but the kinds of changes that may actually occur are not nearly as predictable as they are under the instrumental model.

The second ambiguity in the conceptual model has to do with how groups of policy makers develop consensus regarding the interpretation of evidence. Cohen and Garret (1975) suggest that members of a policy-making group share a variety of assumptions about the problems they hope to solve and the ways in which they are likely to solve them. If this is true, then the collective climate of opinion operates in the same way the individual's preconceived body of knowledge does. It is the thing that must be changed by the evidence, yet it is also the thing that must be applied to the evidence in order to determine its relevance and its meaning. But to the extent that different participants define issues differently, a given piece of evidence may be perceived as more or less relevant to different participants or as equally relevant but differentially instructive.

Collective interpretations of evidence, like individual interpretations, must occur actively rather than passively. Groups must engage in some form of group processes in order to render the evidence meaningful to the group. These group processes can consist of debates, negotiations, bargaining, and other interactions which, then combined, contribute to the formation of a collective point of view which in turn enables a collective interpretation of the evidence.

The variation among participants' perceptions of their substantive terrain and the necessarily correlated variation in their perception of what the evidence means also bears on how the evidence might eventually influence practices, for social services are arranged and delivered by organizations, rather than by individuals, and major changes in services come about only when the collective perceives things differently. Yet there is no a priori reason to believe that all members of the group will change their perceptions in similar directions upon exposure to the evidence. Several possibilities may occur: (a) the evidence may influence some members of the group to change their perceptions but not all members; (b) the evidence may influence all members but influence them each in a unique way; or (c) the evidence may influence all members in a similar way. Although these three possibilities are analytically distinct, and could be hypothesized to have distinct effects on practice, they all fit within the conceptual model of use, at least as we now understand it. But each may have a unique effect on the collective conceptualization of the issues, such that the viewpoints of participants are more unified, more diversified, or as diversified as they were prior to the receipt of the evidence.

The third ambiguity in the conceptual model stems from the fact that, in principle at least, each participant can derive a unique interpretation of the evidence, and each interpretation can be valid from that individual's point of view. One response to this dilemma is to distinguish, as Deising (1962) did, between factual and normative meanings, and suggest that the participants can at least agree on the factual meaning of the evidence. But even facts are not immune to the interpretive process, for their meaning can depend upon where they are placed in the larger structure of knowledge and ideas. No individual can be said to have a greater claim on truth: to some extent, all participants are interpreting the evidence correctly, and to some extent all of them are distorting it in order to fit it into their respective frameworks. And it is not obvious that group processes such as debate, negotiation, and so forth, would necessarily create a higher-fidelity interpretation, for the ultimate consensus may be influenced more by the distribution of manipulative and persuasive skills of the participants than by the evidence per se.

This paper is designed to shed some light on these ambiguities by reviewing seven examples of collective conceptual uses of evidence. The examples were found in a pool of 43 episodes involving the use of evidence in local school districts. These seven examples were selected on the basis of two criteria. First, the evidence had to be received with more than a nod and a smile -- it had to be discussed enough by members of the group to indicate that some form of group processing of the evidence was occurring. Second, the evidence could not have led to a clearly recognizable decision, for if it did, it would have been an example of instrumental use. This is not to say, however, that there could be no effect on practice, for one of the purposes of the review is to determine whether and how conceptually-used evidence can influence practice. Rather, our second criterion merely omits examples in which formal decisions, recognizable as such by participants, occurred.

The episodes came from 16 school districts which participated in a study of school district uses of evaluation and test data. The districts are heterogeneous in such demographic variables as size, wealth, geographic location and the ethnic balance of the student body. The distribution of examples of conceptual uses is not uniform across the districts, but there is nothing in the data to suggest that the distribution is anything other than random. The data gathered in these districts came from observation of groups or from interviews with individuals. The meetings were sampled primarily according to convenience or circumstance, but interviewees were intentionally chosen to include representatives from each level of the hierarchy and, where possible, to include all points of view on a known issue. Rather than asking participants about their uses of evidence, and thereby running the risk of artificially enhancing its use, interviewers asked instead about current affairs, the particular issues with which participants were grappling, and what evidence was available to help them resolve those issues. The episodes described here are about issues which were mentioned by several interviewees and which came up in observed meetings as well.

The seven episodes illustrating collective conceptual uses of evidence are presented here in two groups. The first section presents three episodes in which participants responded to the evidence with diverse interpretations, and the second presents four episodes in which participants, though not completely unified, appeared to have had relatively more success in creating a coherent climate of opinion.

DIVERSE RESPONSES TO EVIDENCE

Three of the episodes in the sample illustrate collective information processing activities under conditions in which individual participants hold diverse points of view. In all three cases, most of the group processes could be referred to as forensic, in that participants either referred to the evidence as they debated the issues themselves, or they debated the meaning of the evidence itself. In none of these examples did we discover any changes in practice, though changes could have occurred after we completed out visits to these districts.

The first episode involved discipline policies and practices in district 18.[1] The issue of whether and when to use corporal punishment came up three years prior to our visits to the district, when some teachers were assaulted by students. Emotions ran high on this issue, and two consecutive superintendents had had problems establishing discipline policies that were acceptable to teachers. The second superintendent's policy -- that students had to misbehave five times before they could be physically punished -- stimulated a general teacher strike, and the strike settlement called both for a survey of the current disciplinary practices and for a joint school district/community task force to review the issues and the evidence and to make recommendations for discipline policy.

Although the issue was ostensibly one of discipline, and was ostensibly one on which teachers differed from administrators, the emotional tensions in the district went beyond this issue. For instance, the district was plagued by a severe distrust between central administrators and teaching staff. Sentiments of this sort tended to arise in all districts participating in this study, but in district 18 these sentiments had evolved from accusations of lack of appreciation for the other point of view to accusations of out and out subversive intentions. These feelings of distrust also extended to the community, so that parents and school staff were squared off against one another. Racial tensions also underlay much of the difficulties in district 18. The mostly white student body had gradually turned to a mostly black student body, while teachers, administrators and even the school board remained primarily white. Interviewees gave several different versions of what the "real" issue was, stating, for instance, that it was really teacher autonomy, or really white teachers who were uncomfortable teaching black students. And they gave different versions of where the polarities really were. Some described it as a disagreement between administrators and teachers, others as an issue between older and younger teachers, and still others as a difference between union-active and non-active teachers.

It was in this atmosphere that the survey of discipline practices was conducted and released to the new Discipline Task Force. This group consisted of administrators, teachers and parents, and the survey findings were presented at its first official meeting one Wednesday evening. The group's first response to the findings was to ask why the response rate was not 100 percent and the evaluator responded by saying, "At some schools the rate was 100 percent but at others it was as low as 30 percent. We had a few mixups in a couple of the schools." As the conversation continued, it veered sharply and frequently, often filling with emotion and often digressing into trivia. But much of it also indicated an interest in trying to understand the nature of student misbehavior and an interest in using evidence to improve their knowledge of the situation. Such interpretive attempts as these were made:

- The problem hasn't improved one iota. You still have 30 percent of the teachers who are only interested in their paychecks.
- Absenteeism is as much a problem among parents as it is among children. We all know what a problem, nationally, absenteeism has become in the workplace. Kids see their parents being absent themselves, so it's the parents who are often setting the example. [This observation was generally agreed upon, and led to a discussion of home discipline strategies.]
- To me the root of the discipline problem is the question of consistency. And that's something that should be enforced from the top down. It's not something that parents can do anything about.

- I think it's even more difficult than that. The underlying causes are in society itself. The problems we face are severe and could take years for this committee to solve.
- [In reference to neighbors who "drink up" their welfare money] Welfare never checks on them neither. So you have the whole system -- parents, teachers, social workers, and everyone -- going around and refusing to accept responsibility. We have this system where no one is paying attention to the other people. So in my opinion, discipline is only part of the problem.
- I still feel we are talking around the issue. . . . As the survey results point out, the problem is not with the policies but with their enforcement. Most of the teachers answered that they don't think the administration is backing them up in discipline cases.
- We know what the top ten problems are. We don't need this survey to tell us. But now that it's done, let's please get on with implementing some kind of strategies.

This group was clearly having a difficult time conceptualizing the discipline issue, and the presence of clear, factual evidence regarding the frequency and distribution of different kinds of student misbehaviors and evidence regarding how teachers were currently handling these problems, did not clarify the issue for them. Though the participants were convinced there was indeed a real problem, they were still at what Rein and White (1977) would call the first stage of problem solving -- that complex process of problem definition, during which groups with competing interests try to interpret the vague and diffuse indicators of stress in their system. Although these participants returned intermittently to the survey findings, they also veered far from them as they discussed and debated their discipline issue. Not only did the group fail to make any major decisions regarding discipline that night, it also failed to determine its own mission -- whether, for instance, it would continue as a district-wide group or divide into a series of independent school-building-based groups, and whether it would first address itself to what is or to what ought to be.

The second episode in which participants approached the evidence from diverse standpoints occurred in district 115. Some time prior to our visits to this district, the state education agency lowered its required coursework in American History, and the district followed suit by changing its configuration of social studies course, replacing its second year of American History with some other requirements. These curriculum changes were not well received by teachers. Many found themselves teaching content areas that did not match their intellectual interests, and some were particularly dissatisfied with one new required course for which no textbook existed. The overall curriculum structure seemed sound to the district, however, and it retained the new design in the hope that the grumblings would dissipate as teachers became accustomed to their new courses.

But two other events occurred before the grumbling dissipated, and these events heightened the visibility of the social studies issue. First, the secondary schools were changed from seven-hour days to six-hour days, thus leaving students with fewer hours in which to take all their required courses. This frustrated students, but they focused their complaints on the rigid sequence of social studies requirements rather than on the reduced length of the school day. The blocks of time actually required for the new social studies sequence are identical to those required under the old curriculum. The second event was that a parent, aware of teacher dissatisfaction, formed a PTA committee to investigate the new curriculum. This group examined patterns of social studies test scores and published its results in the local newspaper. The school district responded to this analysis by conducting one of its own, and the major findings of each analysis -- arguments from the parents and rebuttals from the district -- are presented in Figure 1.

This debate, like district 18's discipline debate, confuses an ostensible issue with several other issues. Many of these other issues were only pertinent to a particular group -- parents, teachers, or students, for instance -- but they complicated things to the point where it was no longer clear what the issue was. Even the director of social studies curriculum vascillated from one point of view to another. At one point, for instance, he said it was good to have people in the community look at the curriculum, while at another point he complained about the parents, saying, "Whenever you start to consider curriculum change you have to remember that we are impacting on 400 teachers, . . . And you have to remember that they are all tenured. The question is, is it worth changing all that just to . . . satisfy 8 parents?" Later he defined the issue as one of raising test scores, saying, "If you want youngsters to do better on test scores, then you just have to move the courses to the eleventh grade, so students don't have a year off before they take the test." And, he added sarcastically, "we should move biology to the eleventh grade because biology scores are down, and we should move math and physics too." Yet at another time he viewed the issue as one of tight scheduling. His own daughter had just moved into

Figure 1: Two Reviews of the American History SAT Scores

Parent-Teacher Association Analysis (Argument)	School District Analysis (Response)
1. The class of 1979, the first class to complete the new curriculum, scored lower in history than in any other subject area.	1. History scores are lower nation-wide than in any other subject area. When these differences are taken into account, local history scores are relatively higher than biology, physics, and math.
2. The class of 1979 scored lower in history than any previous classes in district 115, yet its verbal scores remained stable.	2. National scores are also lower because the test was re-scaled. Taking scale into account, the drop is about 2%, a small price to pay, given the addition of the other subject areas.
3. District 115 scored lower than the district next door, yet in previous years district 115 scored higher than the neighbor.	3. Only 70 of the neighboring students took the test, whereas 300 of our students took it. Furthermore, their students take history in eleventh and twelfth grades, whereas our students take it in ninth and tenth grades.

high school, and: "I can see the problem more clearly now. It's going on right in her school. . . . With the history and social studies requirement, they don't have time to take all the courses they want."

Building staff, on the other hand, seemed not to know that the issue had even begun with them. They assumed that the issue had to do with the quality of their social studies program, rather than with the particular courses that were required, and they also erroneously assumed that the flap over test scores was centered on the standardized, norm-referenced achievement test given annually to eleventh graders, rather than on the scholastic aptitude tests which were voluntarily taken by college-oriented students. They defended their social studies program by pointing out that the test content was not matched to their curriculum and never had been. The test carried items about ancient history, which the district did not offer, and items on modern history, which the district did not cover. Although the emotions in district 115 did not run as high as they did in district 18, district 115's conceptual struggle was imbued with a similar confusion as people moved between evidence and a tangle of issues in an attempt to clarify "the" issue.

The third example involving diverse points of view also arose in district 115. It involved evidence to the effect that the proportion of minority children enrolled in the special education classrooms was larger than the district's overall proportion of minority students. This evidence was legally damaging, since the district could have been sued by the United States Office for Civil Rights on the basis of these enrollment disproportions. Consequently the evidence was suppressed, even though the administration continued to struggle with it. The issues surrounding the debate included such things as the procedures by which children were referred to and assessed for special education, the training and qualifications of assessment personnel, the degree to which individual placement decisions were reviewed by people higher up in the system, the availability of classrooms for children who were not retarded but who nonetheless were behind academically, and so forth.

These three episodes share several features in addition to the fact that the evidence was interpreted in a variety of ways. One feature they share is that the issues were not clearly defined. In all cases, they were confused by the introduction of tangential issues and digressions. Such a muddling is a natural outgrowth of the fact that conceptual use is an individualized activity before it is a group activity. As each individual encorporates the evidence into his or her own framework of knowledge, each associates it with a unique set of issues -- for one participant, the data are relevant to scheduling, and for another they are relevant to staff training and oversight. Each additional participant in the discussion expands the scope of the issue by introducing new peripheral issues, until members are confused not only about what solutions should be considered, but about what the problem itself even is. Rather than clarifying issues, then, the evidence stimulates a debate which goes far beyond the evidence itself, to a wide range of often only loosely connected issues.

The other important feature which these episodes share is that they did not appear to result in changes in practice. It seems reasonable to suppose that no changes could occur until the group at least agreed on what the issues were and what those issues implied for changing practices.

These groups could not agree on that point, and their group processes were aimed more at developing a unified point of view regarding the substantive terrain than at defining any particular changes in practice.

ATTEMPTS AT UNIFYING RESPONSES TO THE EVIDENCE

Even groups who are in relative agreement on most things are not in perfect agreement, and members of organizations spend a great deal of time trading ideas and checking their perceptions with one another. These interactive processes enable individuals to modify the consensus by inserting new ideas, as well as to modify their own ideas to better adapt to the consensus. When new evidence enters an organization, it is treated much the same as new ideas from any individuals within the organization. It can modify the prevailing point of view, but the prevailing point of view can also modify the meaning of the evidence. These changes can come about so gradually and subtly as to be almost unnoticed. The four episodes reviewed in this section differ from the first three in that participants in these episodes were not vigorously debating either the issues or the evidence. Though their points of view were not precisely the same, they at least suggested that broad agreement had been reached.

The first episode occurred in district 57, and involved the use of teachers' aides. District 57 had several categorical programs -- bilingual programs, compensatory education programs, and so forth -- which employed aides. Everyone had assumed that the use of aides was a programmatically sound idea; their presence was never questions. But because they constituted a substantial portion of the budget, they eventually came under the scrutiny of the evaluation unit, and it conducted a study which indicated that the presence of aides did not have a significant effect on children's achievement test scores. This finding surprised nearly everyone, and in fact was simply not believed at first. Even the evaluator was surprised. After reading about the benefits of aides in other programs around the country, he fully expected his study to demonstrate positive effects. He repeated the study the following years, and has since intermittently repeated it in various programs, with consistent results.

The reaction to these studies was unified in the sense that participants did not believe the evidence at first, and in the sense that many of them became convinced after the study was repeated. It was also unified in the sense that everyone described the findings using the same phrase -- not that aides were ineffective, but that they did not make a difference in achievement. But the response was not unified in the sense that all members agreed on exactly what changes should have been made, not even on what changes had in fact been made. References to changes that came about as a result of the study included these: [Emphasis is added to each quote.]

- (Superintendent) Spreading the decision out over two years [because the findings were not believed at first] made it easier to phase out aides.
- (Board member) As I recall, it was a formal decision and was even voted on as policy.
- (Associate Superintendent) We were inclined to cut back on the number of aides working in these programs. There were no across-the-board cuts -- instead, the study influenced the approval of aides in new programs and in the review of old programs as they came up.

- (Program Director) I have made an effort to cut aides from my programs.
- (Program Director) The board is now resistent to proposals that include aides.
- (Principal) I think people have tended to try and do away with aides since the study came out.
- (Teacher) I lost my aide last year because aides were pretty much terminated in the program. [This teacher was unaware that any studies on this topic had been done, and perceived this as an entirely a programmatic decision.] But next year I'll get my aide back again because the program will be run by [someone else] who thinks an aide will be helpful. [Not having an aide last year] was a great loss to us here because it was important for the students to have as much contact with adults as possible.
- (Teacher) [pointing to two aides working with her children as she talks with us] I am aware that kids are better off spending time with teachers than with aides, and I do my best to spend a lot of time with the children myself, rather than letting the aides do all the work.

Though one person indicated that a clear decision, even a vote, had been made, everyone else suggested that they had shifted their posture toward aides in ways that would ultimately either reduce the number of aides employed or change the duties assigned to them. The conceptual effect of the study was to cause them to scrutinize more closely a host of day-to-day decisions that had in the past been made without question. But no one could define in concrete terms exactly what the district's agreed-upon response to the evidence actually was. The response was unified only at a very general level.

The second episode also occurred in district 57. In this case, the study was a descriptive study of how classroom time was allocated in compensatory education classrooms as opposed to regular education classrooms. Among other things, it indicated that students in compensatory education received less instructional time in science and social studies than students in regular education did. As this message traveled about the district, it became abbreviated. Several members described it not as a differential shortage, but instead as a universal shortage. For instance, a compensatory education teacher joined regular education teachers who were searching for ways to increase social science instructional time. As she described their deliberations, she said, "And we even found a way that compensatory education can make a contribution. For example, I can teach map reading . . . " [emphasis added]. By the time the message had reached school buildings, then, it was stimulating participants to increase time spent on social studies in regular education rather than specifically in compensatory education, where the shortage had been found.

The third episode illustrates a similar process of gradual distortion as various participants attempt to digest the evidence and determine its meaning. This episode occurred in district 220. Like many other school districts, district 220 had vascillated over the past two decades in the amount of centralized control it exerted over the curriculum. In the early seventies, the reading curriculum was changed from centralized to decentralized, and the superintendent had encouraged building principals to be creative and independent. The resulting diversity in the reading curriculum and in the instructional practices used to teach reading set the stage for a study of the relationship among these various practices and improvement in reading achievement, and eventually to a new interest in centralized control of reading.

The study correlated a large number of variables with gains in reading achievement, and the final report listed variables that were found not to be related to reading growth as well as those that were. For instance, it noted that certain characteristics of building principals were not related to growth in reading scores: their administrative experience, their experience in the particular school, and whether or not they held advanced degrees. Most of the report was focused on variables that were related to reading achievement, of course, and the report included many recommendations based on these findings. For instance, children who had attended kindergarten classes were found to gain more than other children, and the report recommended that the district try to increase enrollment in kindergarten classes; children using a particular text gained more than other children and the report suggested that the district explore ways to increase the use of that reading series; and children gained more when their building principals observed classroom sessions, and the report urged an increase in the time principals spent observing classrooms.

The study received a great deal of attention when it was released, and the superintendent then appointed a committee to develop a major new Achievement Plan for the district. The reading correlates study was to be one of the committee's primary references. After six months of deliberation, the committee presented its recommendations to the district. Within its 35 recommendations were seven references to the Correlates to Reading Improvement Study and three recommendations that flowed directly from it. One of the study's findings, that principals' observations of classrooms were correlated with gains in reading, was wrapped in an elaborate interpretive package as the committee report referred to the study as having found that achievement went up "when principals took a direct, active role in putting together the reading program and spent a good portion of their time in classroom monitoring." This interpretation of the evidence led the committee to recommend that principals be provided with leadership training. Later on, the report further suggested that, if reading scores went up when principals were actively involved in the "concept, implementation, coordination and evaluation" of reading, then it stands to reason that the district superintendent should be just as involved in the reading plans for individual schools. Whereas the original research report had focused on curriculum, pedagogy, or other aspects of direct service, the committee's report focused on administrative and managerial issues -- procedures for oversight, planning, and coordination.

The Achievement Plan received as much attention as the study itself had six months earlier. Both were covered in the local newspapers and both made the agenda of the superintendent's cabinet on more than one occasion. As a result, many members of the district referred to one or the other of these documents, and often their references indicated a misunderstanding regarding what the research findings actually were. For instance, a program director, when discussing his plans for allocating program funds, said, "I met with [an associate superintendent] to discuss the most critical needs. The first level he identified was staff training, especially at the principal level. This was supported by the Correlates to Reading Improvement Study." Another program director referred to the Achievement Plan as if it had no known basis in evidence, saying, "How do they know what works? That's the key role for [the evaluator] and his people. I'm going to ask [the evaluator] to address the Achievement Plan Committee." Finally, we met a principal who referred to the study as having found that "there are basically two types of principals, the public relations type and the curriculum and instruction type." This principal was especially proud of the fact that he was the curriculum and instruction type of principal. And he was proud of the fact that he never monitored instruction by observing classrooms, but instead relied solely on his reviews of children's test performance.

The fourth episode took place in district 115. This district, in an effort to compare its community's perception of education with nationwide perceptions, conducted a local opinion poll using items from the Gallup Poll on attitudes toward education. The findings were presented in a question-and-answer brochure, with data on the respondents' characteristics tucked in the back. The findings were also reported orally to the superintendent's cabinet, and the oral presentation played a particularly important role in the interpretation of the evidence. During that conversation, a demographic statement that might have otherwise remained buried in the back leapt forward and captured the attention of the group, at the expense of all the other findings: only 30 percent of the respondents had children in the schools. That the majority of citizens did not have school-aged children was apparently a surprise to cabinet members, and although none mentioned the survey's findings about attitudes when they were privately interviewed, three volunteered the demographic finding, and each distorted it slightly in repeating it to us:

- o For the first time, non-parents are now the majority of the population in district 115. [Emphasis added]
- o For the first time, we have more of our citizens who are dealing with the schools who don't have school-aged children. [Where did you get this fact?] It's an acknowledged fact, but I can't put my finger on where I learned it. [Emphasis added]
- o Over 51 percent of our citizens in district 115 do not have children in the system.

These four episodes illustrate the effect of group processes on the interpretation of evidence. They are similar to the first three episodes, for in both sets of episodes participants appeared to be trying to establish an agreed-upon interpretation of the evidence. But those who were involved in the first three episodes were unable to do so because their diversity drew too many peripheral issues into the discussion. Those involved in the latter episodes had achieved at least a broad consensus.

But the consensus attained in these latter episodes was only tangentially related to the evidence. A discrete correlation between principals' observations and reading scores became a pronouncement about the importance of principals, which each participant could then elaborate with his or her own examples of important principal behaviors. A particular finding regarding differences in instructional time between two programs was raised to a general statement about the importance of spending more time overall. Findings about the community were changed from new knowledge about the community to a new phenomenon in the community. In each case, the original message was changed as the group processes it. It was translated to a slogan that everyone could agree with and which could imply a general direction for change in practice, with little regard for how faithful the message was to the original evidence.

These latter episodes also differ from earlier episodes in that three of them did in fact lead to changes in practice, whereas none on the first three led to such changes. Although our sample sizes are too small to warrant inferences regarding this apparent relationship between diversity of viewpoints and changes in practice, the properties of the episodes themselves suggest that there may indeed be a relationship. When participants respond diversely to the evidence their debates tend not to be unidimensional debates, but instead are jumbled arguments about a host of loosely connected issues. The tangle is so complete that agreement is not possible even at a very general level. Changes did occur, on the other hand, when participants could agree to modify their positions on a broad class of events and practices, so that each could change a variety of decisions and actions in the future. The agreement was not so specific, though, that any one participant could predict what others were doing differently. And the agreement was only met by modifying the message presumed to have come from the evidence.

CONCLUSION

Our purpose has been to explore the implications of the conceptual model of how evidence is used. Of particular interest were three aspects of the model that have not been clearly defined as it now stands. One of these is the relationship between conceptual uses and changes in practice. The second is how groups process information to create a collective interpretation, and the third is the question of fidelity in interpretation when conceptual use occurs. We suspected that these questions might be related. Even when the collective does not make crisp decisions by votes or other means, it still engages in a variety of group processes which establish a climate of opinion. That climate, in turn, may be more influenced by the distribution of manipulative and persuasive skills among participants than by the facts themselves, but it would nevertheless be used by individuals to judge the appropriateness of their diverse decisions and actions.

Upon reviewing three episodes in which members of the school districts had diverse reactions to the evidence and four others in which members established unified interpretations, we find support for these contentions. Changes in practice only occurred in those cases when members agreed at least on one or two slogans that were implied by the evidence. However, these examples also suggest that such agreements came about only when the group took considerable liberties in its interpretations of the evidence, and some of the resulting distortions were considerable. The whole notion of conceptual use, then, appears to rest on a Catch 22: without some degree of agreement, no operating principles can be inferred; yet agreement regarding the meaning of the evidence can only be generated by giving the evidence a meaning that participants can agree on.

NOTES

[1] School district code numbers indicate the approximate size of the district in thousands of students served. The smallest district, district 4, serves roughly 4000 students, and the largest, district 240, serves over 200,000 students. Actual service rates randomly vary by ±15 percent from the rate implied by the code.

REFERENCES

Allison, G. Essence of Decision: Explaining the Cuban Missile Crisis. Boston: Little, Brown, 1971.

Alkin, M., Daillak, R., and White, P. Using Evaluations: Does Evaluation make a Difference? Beverly Hills: Sage, 1979.

Caplan, N. A Minimal Set of Conditions Necessary for the Utilization of Social Science Knowledge in Policy Formation at the National Level. In C. H. Weiss (Ed.), Using Social Research in Public Policy Making. Lexington, MA: D. C. Heath, 1977.

Cohen, D. K. and Garret, M. S. Reforming educational policy with applied research. Harvard Educational Review, 1975, 45, 17-43.

Cohen, D. K. and Weiss, J. Social Science and Social Policy: Schools and Race. The Educational Forum, 1977, May, 393-413.

Deising, P. Reason in Society: Five Types of Decisions and their Social Conditions. Urbana, IL: The University of Illinois Press, 1962.

Leviton, L. C. and Hughes, F. X. Research on the Utilization of Evaluations: A Review and Synthesis. Evaluation Review, 1981, 5, 525-548.

Rein, M. and White, S. Policy Research: Belief and Doubt. Policy Analysis, 1977, 3(2), 239-271.

Rich, R. F. Use of Social Science Information by Federal Bureaucrats: Knowledge for Action versus Knowledge for Understanding. In C. H. Weiss (Ed.), Using Social Research for Public Policy Making. Lexington, MA: D. C. Heath, 1977.

Weiss, C. H. Research for Policy's Sake: The Enlightenment Function of Social Research. Policy Analysis, 1977, 3(4), 531-545.

Weiss, C. H. Knowledge Creep and Decision Accretion. Knowledge: Creation, Diffusion, Utilization, 1980, 1(3), 381-404.

EVIDENCE AND DECISION

The two concepts mentioned in the title of this paper -- evidence and decision -- are often assumed to hold a special relationship. In its simplest form, a decision is a choice among two or more options and evidence is the stuff that informs the choice. And to the extent that the choice is based on evidence, particularly evidence about the costs and benefits of each option, the decision is considered to be sound, or rational. Of course, not all decisions are as simple as this, but even complex decisions are assumed to be made better if they are based on some form of evidence. This kind of decision making is often called technical, or scientific rationality, because of its reliance on technical or scientific evidence. Belief in the rightness of rational, evidence-based decision making has created a demand for the kinds of evidence deemed necessary for rational public decisions -- applied social science, program evaluation, management information systems, and policy analyses.

But studies of the decision-making process itself have suggested that these desireable characteristics rarely occur in practice. The language now used to describe decision making is filled with such terms as "satisficing" (Simon, 1957), "mutual adjustments" (Lindblom, 1965), and even "garbage cans" (March and Olsen, 1979), rather than such adjectives for rationality as "efficiency", "goal attainment", and "maximizing utility". These observations indicate that the ideal form of rational decision-making rarely occurs. Instead of a clean crisp line between evidence and decision, there are multiple wavering lines and several smudges as well.

These observations have not been greeted with complete dismay, for they have been accompanied by a recognition that value judgments can also make an important contribution to decision making, that value judgments can, on occasion, legitimately usurp the evidence, and that those judgments should be expressed, clarified, and debated by means of participatory decision-making. It is possible, in other words, for a decision-making process to be a good one even if the resulting decision does not clearly conform to the evidence. Still, there is an uncomfortable disjuncture between the requirements of technical rationality and the requirements of participatory decision making. And although observers acknowledge the value of participatory decision making, they are not willing to completely abandon their belief in the rightness of scientific rationality. Consequently, some authors have tried to re-allign these two disparate ideals to create a compromise model of rationality.

One of these authors, Paul Deising (1962), has argued that there are several kinds of rationality -- technical, economic, social, legal, and political -- and that some contexts may require balancing all of these kinds of reasoning. Consider, for instance, the several paths by which an educational decision could be reached. To the extent that education is considered to be an applied science, then the standards of the rational model of decision making could apply to educational reasoning. There does exist, for instance, a body of research on pedagogical and programmatic practices, and most large school districts also maintain management data such as enrollment or attendance statistics and students' test records, all of which could provide the technical basis for rational educational decisions. Yet there are also times when technical rationality may seem less appropriate than, say, legal reasoning regarding the rights of students or the rights of employees. The body of legal decisions that bear on educational practice has grown considerably in recent years and now embodies many substantive areas that might formerly have been considered technical. For instance, a handicapped child who in the past might have been placed in a special education program on the basis of research regarding the effectiveness of that program might today be retained in the regular classroom on the grounds that his or her right to full participation is being abridged by removal from that environment. In still other circumstances, both legal and technical reasoning may be abandoned in favor of political reasoning: the care and education of children is an extremely value-laden enterprise, and parents have different and often extremely diverse ideas about how their own and other children should be treated. The more vocal of them are continually pressing their points of view on school district decision makers. Finally, Deising also argues that social interactions are guided by a social form of rationality: relationships are characterized by reciprocal or complimentary sets of expectations and obligations, and people strive for parsimony and balance in these relationships. To the extent that social relationships among educational decision makers influence their decisions, then, the substantive resolution of an educational issue may simultaneously involve several kinds of "rationality."

Of the different kinds of reasoning that could be used, Deising argues that political reasoning will always take precedence over the others. And for Deising, a "good" political decision is one which, on the one hand, enables diverse ideas and options to be introduced, and yet on the other enables a unified resolution to be developed. Since these two standards pull in opposite directions, the problem inherent in developing or maintaining a sound decision-making structure is one of striking a balance between these two requirements. If diversity is too great, the group may find consensus impossible, and may eventually fall apart. On the other hand, if the group is too cohesive, no real diversity of opinion is available, so no real choice is made. Deising's compromise version of rationality emphasizes participation far more than evidence. As long as the optimal balance between diversity and unity is met, Deising is not very concerned about whether evidence plays an instrumental role in the development of the unified resolution. Instead, he expects its value to be primarily one of stimulating ideas which, in turn, are modified or combined by means of group processes. For Deising, then, what matters is not the particular role that evidence plays, but rather whether the political processes enable a unified resolution to be carved from diverse opinions and options.

A second author, Yaron Ezrahi (1980), approaches the problem of compromise by distinguishing between "utopian rationality" and "pragmatic rationality," the former being the kind of decision making in which there is a clear, unwavering relationship between evidence and decision, and the latter encorporating participatory processes as well as evidence. For Ezrahi, who is a pragmatic rationalist, good decision making cannot occur when decision makers replace scientific standards with political standards, nor when they do the reverse. Instead, they must "fuse knowledge and policy within the limits of political and moral requirements and by the standards of scientific truth and rationality" (Ezrahi, 1980, p. 131, emphasis added). While Deising has suggested that evidence may play its most important role in stimulating ideas, Ezrahi has suggested that there still is an instrumental use for evidence. But he does not discuss exactly how these two sets of standards -- political and moral requirements on the one hand and scientific truth and rationality on the other -- are to be simultaneously maintained. For instance, he does not consider the possibility that evidence may be misused in a highly emotional debate.

A third attempt at compromise is offered by Cook, Levinson-Rose and Pollard (1980), who are more interested than either of the other authors in maintaining the standards of scientific rationality. While acknowledging the inevitability and value of participatory decision making, their aim is to find standards for appropriate use of evidence within a political decision-making context. Their "normative model" of the use of evidence is one in which evaluation results are considered to be "inputs into a debate, and that is all" (Cook et al., 1980, p. 482). Under the old ideal of rational decision-making, which these authors call naive, proper use of evidence was gauged according to whether the option actually chosen was the one favored by the evidence. Lack of concurrence between evidence and decision was frowned upon as not rational. Recognizing that concurrence is too stringent a criterion, and that it completely precludes political activity, these authors propose some new standards by which to judge how well the evidence has been used. Their standards include the extent to which the evidence is (a) accurate, (b) accurately disseminated, (c) disseminated to all groups or individuals with an interest in the issue or decision, (d) interpreted without bias, and (e) used in a fair debate. In an attempt to develop a new list of standards for judging how well evidence has been used in a political climate, these authors have been more specific than either Deising or Ezrahi, but have also indicated a greater preference for the scientific standards than have the other authors.

All of these authors are interested in developing a compromise model of rationality, one that can encorporate the ideals of both participatory and evidence-based decision making. Though they differ in their relative emphasis on each side of the equation -- Deising emphasizing participation, Ezrahi trying for a balance and Cook and his colleagues emphasizing the scientific -- they are nonetheless similar in two important respects. First, all of them recognize that the diversity of ideas and opinions from which decisions flow encompasses more than simply technical differences among strategies; that it also encorporates differences in assumptions about purposes and about the legitimacy of various means-ends relationships. Second, all these authors recognize the latter diversity as legitimate, while still assuming that evidence also has a legitimate role. The old model of rational decision making, called "technical" by Deising, "utopian" by Ezrahi, and "naive" by Cook and his colleagues, is simply too limited to be useful. All recognize the limitations of the old concept of rationality and all are interested in preserving the spirit of it -- they would rather redefine the term than abandon it. Consequently, all have tried to merge what had previously been two different sets of decision-making standards -- political and scientific -- into a common framework that could be considered "rational."

Broadly speaking, these proposals suggest that a reasonable compromise may be possible, but none is so well developed that it defines the specific way in which evidence should be used in participatory decision making. Two questions are particularly important. One has to do with the process of developing a unified resolution from diverse ideas and options; that is, how can evidence facilitate the processes by which points of view are unified? The other has to do with whether and how evidence influences the substance of the eventual decision; that is, under what circumstances can or should the eventual decision concur with the evidence?

The political processes used to develop unified resolutions include such activities as discussion, debate, brainstorming, negotiation, bargaining, and compromise, all of which are clearly different from dispassionate and rational reliance on evidence. Some of these processes, while considered legitimate politically, may not be considered legitimate even under a compromise model of rationality. For instance, instead of expressing their diverse points of view in open debate, participants may choose more hidden methods for checking the spread of their opponents' ideas, or they may intentionally shift a debate from, say, technical to bureaucratic issues, simply in order to jam the unification process. Furthermore, these processes are not always public, so participants may not always feel accountable to that vagely defined audience which is usually assumed to require at least the appearance of rationality. What is not clear in the compromise models is whether evidence is merely one of many tools, used well or poorly by diverse participants as they bargain, negotiate, brainstorm, and so forth, or whether the evidence should still have an independent effect on the diverse points of view, such that it facilitates unification apart from its use as a tool in these political processes.

With regard to the option actually chosen, the old model of rationality assumed that the evidence would clarify the relationship between means and ends with a logic so compelling that decision makers would have no recourse but to choose the option indicated by the data. Furthermore, under the old model, the evidence was able to serve this rational purpose in large part because of the methods by which it was generated -- methods that endowed it with an aura of indisputability. The compromise models do not compel decision makers to act on the basis of the evidence, and in fact they distinguish the factual meaning of the evidence from its normative meaning, thereby acknowledging that it may very well not be compelling. When decision makers have widely divergent points of view, it is possible that they will not even find the same evidence to be relevant, let alone finding it to be instructive. Yet if one acknowledges that evidence can have multiple meanings, then one introduces a host of questions about the difference between its facual meaning and its normative message, and about its potential to make a substantive contribution to the forthcoming decision.

The compromise models, then, raise two important questions regarding the use of evidence. First, can it have an independent effect on the otherwise purely political process of consensus-building, or is it merely one of many tools, used both well and poorly, to carry out these political processes? Second, under what circumstances can the substance of the eventual decision be expected to concur with the evidence? In fact, these two questions are related. For to the extent that the evidence has an independent influence, that influence must rest on a clear factual basis that compels participants to favor a particular option; and when it lacks independent influence and is instead merely a tool used well or badly by all participants, then whatever unified resolution develops will not be determined by the substance of the evidence but by the distribution of political and manipulative skills among participants.

If evidence never had an independent unifying effect of its own, there could still exist many decisions which agreed with the substance of the evidence that was brought to them. But one would expect the substantive correspondence between evidence and decision in a population of decisions to be randomly distributed. We therefore cannot assume that occasional substantive concurrence between evidence and decision necessarily indicates that the evidence has had an independent effect, that the decision-making process has met the standards of scientific rationality, or even that it has met the less-well-defined standards of compromised rationality, for such a pattern may reflect no more than the coincidental outcome of an entirely political process. One proxy, however, that may be useful in estimating the influence of evidence, relative to political influences, is the temporal relationship between the evidence and the decision. If the decision follows relatively quickly upon the release of the evidence, it may be because the evidence was sufficiently compelling to have a unifying effect on participants. If, on the other hand, a decision lags considerably behind the evidence, it may be because the evidence did not unify points of view, and time was needed for the political, social, or organizational processes to do that job. There are also, of course, cases in which the decision precedes the release of the evidence. Presumably these cases would indicate that the unification was accomplished entirely by non-evidential methods.

This paper is designed to shed light on the compromise model of rationality by reviewing the details of 14 decisions which involve evidence that at least some participants claimed was instructive to the decision. The episodes came from school districts which participated in a study of the uses of evaluation and test information. The sample of districts is heterogeneous on such demographic variables as geographic location, size, wealth, and ethnic balance of the student body. The distribution of the episodes across these districts is not uniform, but there is nothing to suggest that they are anything other than randomly distributed. The data gathered in these districts came either from observations of group meetings or from interviews with individuals. Observed meetings and individual interviewees were chosen primarily by convenience of circumstances, although interviewees were also chosen to include representatives from each level of the hierarchy and, where possible, to include all points of view on a known issue. Rather than asking participants about their use of evidence per se, and perhaps thereby inviting them to invent uses, interviewers asked participants about current events, the particular issues with which they were grappling, and what evidence was available to help them resolve these issues. In this way, the role of the evidence was not artificially enhanced.

The next section of the paper describes the decision-making context of school districts. That section is followed by a review of the 14 decisions that occurred in these particular districts, and then by a concluding section which reviews the contribution evidence appears to have made to these decisions.

THE DECISION-MAKING CONTEXT

School districts are the primary administrative units through which educational services are delivered. Like fire departments or police departments, they are a part of local government. Their existence is authorized by state laws, and their authority and mission is prescribed by state law. But their budgets and many of their policies are connected to city, township, or county government budgets and policies. There are nearly 16,000 school districts in the country, and the majority of them are quite small -- nearly 30 percent serve fewer than 300 children, and only about 45 percent have more than 1,000 pupils (National Center for Educational Statistics, 1980).

Generally speaking, school districts have two distinct parts. The first and larger part is the administrative unit which actually organizes and provides services. The second and politically more important part is the school board, which usually consists of from five to nine members who may be either elected or appointed through local governmental processes, and which determines the district's budget and policies. The distinction between issues that are primarily administrative, and hence the province only of the administration, and those that involve budget or policy, and hence require participation of the board, is not at all clear, and districts differ considerably in the range of issues with which their boards deal.

For most of the larger school districts, regardless of the variety of reasoning processes that may be involved -- legal, political, or scientific, for instance -- the decision-making structure is primarily bureaucratic, and the decision-making unit is usually a small group of people which has been assigned the decision-making task. One of these groups, of course, is the school board, a group that may meet anywhere from biweekly to monthly, depending on how involved it becomes in various issues. The agenda for these meetings is usually determined by negotiation between the superintendent and the school board, though the superintendent is usually primarily responsible for it. Because the board establishes official policies, its decisions are usually made by voting. Another common group is one convened by the superintendent, consisting of senior administrators, which usually meets weekly. In relatively small school districts the superintendent's cabinet may consist either of building principals or of two or three close associates who have, by virtue of their personal relationship with the superintendent, been given administrative titles and responsibilities that justify the superintendent's reliance on them. In larger districts, the cabinet is more likely to consist of people whose titles are deputy, assistant, or associate superintendent. If a superintendent is unable to change the people who hold these titles because of employment rules, he may choose instead either to redefine the composition of the cabinet or to stop convening it. The cabinets in the districts involved in this study ranged in size from three members to over thirty. They rarely voted. Instead, they either reached informal consensus or vigorously debated issues so that the superintendent, who would ultimately make the decision, would have a chance to hear all sides of the issue.

These two groups, the superintendent's cabinet and the school board, constitute the funnel through which all major decisions must pass, and usually through which most ideas intended to influence other points of view must also pass. The ideas that enter these groups come from several sources. Board members, for instance, receive ideas through informal conversations with neighbors, attendance at various civic functions, letters, and board hearings. Members of the superintendent's cabinet are all supervisors of portions of the district staff, and receive ideas from their staff. Furthermore, most districts have a number of other officially-recognized groups -- committees and task forces of various kinds -- which eventually report through the cabinet and to the board.

The volume of conversation that can occur in a school district is partially indicated by the number of these committees, and even relatively small school districts can have a substantial amount of committee activity. Figure 1 sketches the committee structure in one of the smaller districts in this study, district 7.[1] Each of the groups listed in Figure 1 meets regularly and in between meetings members discuss issues with their colleagues, so that there is a constant flow of ideas passing between committee members and their colleagues as well as among the committee members themselves. The development of an analogous "map" of the committee structure in a larger district would be a monumental undertaking, though its presence is indicated in several ways. For instance, in district 220, the director of mathematics curriculum explained his communication network by listing a host of committees: a mathematics advisory committee that looks at the math curriculum across the entire grade span, an elementary math committee that consists of elementary teachers and principals and junior and senior high committees that are comprised of the department chairpersons at each of these levels. Most

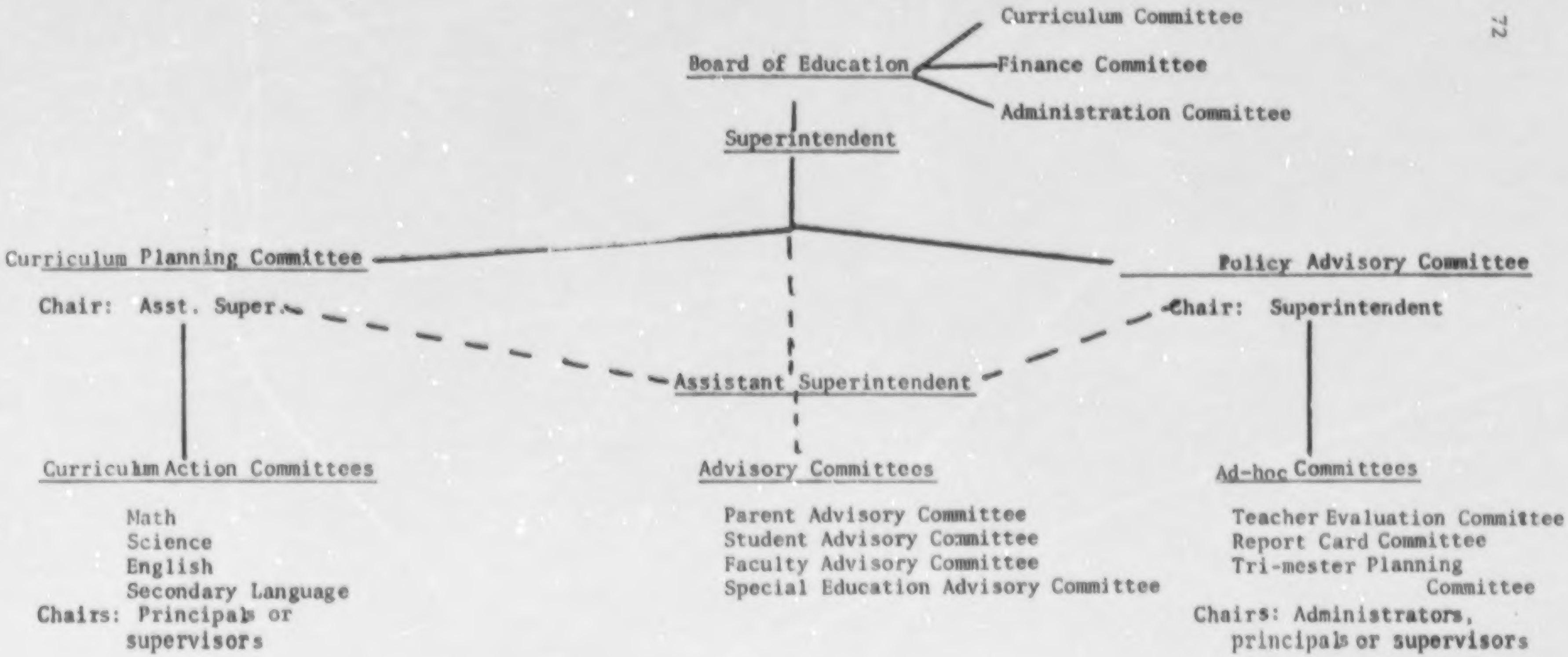
Board of Education
Curriculum Committee
Finance Committee
Administration Committee
Superintendent
Curriculum Planning Committee
Chair: Asst. Super.
Policy Advisory Committee
Chair: Superintendent
Assistant Superintendent
Curriculum Action Committees
Math
Science
English
Secondary Language
Chairs: Principals or supervisors
Advisory Committees
Parent Advisory Committee
Student Advisory Committee
Faculty Advisory Committee
Special Education Advisory Committee
Ad-hoc Committees
Teacher Evaluation Committee
Report Card Committee
Tri-mester Planning Committee
Chairs: Administrators, principals or supervisors

FIGURE 1: The committee structure in school district 7

of these committees meet monthly, and our director of math tries to attend all of these meetings. Furthermore, he meets regularly with two groups of resource teachers, people trained especially to help other teachers understand and apply the district's math curriculum, with parents to interpret the curriculum for them, and with area superintendents (this district is sufficiently large that it is subdivided into regions, each with its own area superintendent), and with school building principals. And, of course, he meets regularly with his own immediate staff.

Although many of these committees serve primarily to coordinate, many are also designed to contribute to substantive policy, and these committees tend to send formal reports up the hierarchy to the superintendent's cabinet, which in turn reports through the superintendent to the school board. The board then responds and the message runs back down the same series of committees. The sequence of events associated with a drug abuse problem in one school district (district 115) illustrates the process. The superintendent authorized a task force of students and local police to study the substance abuse problem and to make recommendations to the district. This task force met for about a year, and then reported to the superintendent with 31 recommendations. Then the superintendent appointed another task force consisting entirely of school district staff, to respond to these recommendations. According to the chairman of the second task force, this group "agreed, disagreed, and reacted to each of the 31 recommendations." This second group's response was then sent to the superintendent's cabinet for review, and with his cabinet's guidance, the superintendent made a series of recommendations to the school board. Among his recommendations were several for the establishment of still more committees: elementary, junior high, and senior high committees, and an interagency committee whose purpose would be to set up a major conference in the Spring. Each of these new committees was to take a look at what the needs were in each of the areas where recommendations had been made. As an example of what these several new committees were doing at the time of our field work, one assistant superintendent described a committee that had been addressing suspension problems. The committee had recently brought recommendations back to the superintendent's cabinet, but the cabinet felt that the timing was not right to take these on to the board. So the ideas were returned to the committee for more work.

In addition to elaborate committee structures which are superimposed over the hierarchical organizations of school districts, there are also formalized procedures for communicating up and down the hierarchy. For instance, many of these districts have annual goal-setting procedures, in which each supervisor-subordinate dyad must sit together and negotiate the subordinate's goals for the coming year. These agreements are usually written down, but districts vary considerably in how formal the agreements are, the extent to which they must be reviewed and approved by higher levels of the bureaucracy, and the extent to which these goals, once established, are ever mentioned again during the remainder of the year. Like committees, though, these procedures increase the volume of conversation that occurs in school districts.

Both of these practices can contribute to political methods of decision-making in the sense that they can facilitate the development of a unified point of view. Messages and countermessages continuously flow up and down and round and round, making it possible for committees as well as individuals to adapt to shifting moods and to modify their own thinking to account for prevailing points of view. For some accommodation may mean changing one's mind, while for others, it may mean developing a new set of arguments for pursuing their old interests.

The effects of these complex networks on points of view can be extremely subtle. For instance, we found several interviewees in one district who had no obvious reason to communicate directly with one another, nor even to be acquainted, yet who nevertheless used a common argument and even a common phrase -- "too many variables" -- to discuss a variety of topics of concern to them in their particular duties. The argument was offered, for instance, by the district's director of the testing program to indicate why he couldn't predict who was likely to use test data and who was not. It was also offered by an elementary school principal to explain why he would not change his school building's practices on the basis of one year's test data -- "too many variables involved to make a judgment on just one year," -- and by a teacher in a different school to explain why teachers shouldn't be evaluated on the basis of children's tested achievement: "there are so many variables involved in working on a human product." And by a program director as a reason why programs should not be required to stand or fall on the basis of a single evaluation. Other occasions in which this argument came up were organizationally related to the individuals cited here, but the examples cited here came from people who had no opportunities to communicate directly with one another. The argument expressed by them had apparently permeated the district and become a unifying theme, a part of each individual's point of view.

THE DECISIONS

Decisions can involve a lot of people and take a lot of time, and circumstances can change even as the decision is being made. Adding to that, the substantive terrain of a decision may be difficult to fully describe, for each can have its own mood, as well as its own array of particular topographical features. Of the 14 decisions of interest here, each has its own educational subject matter, of course, but most also entail matters of economics, politics, organization, or individual personalities. All of the decisions are plagued with some sort of complications, and require elaborate description in order to illustrate exactly how the evidence has entered into them. Such portrayals would, unfortunately, be too lengthy for these pages. Rather than compromise the length of them all, we present some in detail and gloss over others, providing the rationale for each presentation as we encounter it. The episodes are presented here according to the temporal relationship between the evidence and the decision. Of the 14 decisions, five seem to have been made before the evidence was released, three were made some time after the evidence was released, and the remaining six were made immediately after the evidence was released.

Foregone Conclusions

The five decisions grouped together here are in fact of two types, but they share one common feature: the decisions appear to have been made before the relevant evidence was available. In three cases, a pilot program was taken to be effective and was expanded before the evaluative evidence was available. In the other two cases, studies were initiated to determine the effects of decisions that had already been made, but the atmosphere surrounding those decisions was such that the evidence was very unlikely to overturn the decision.

Let us first consider the two cases in which the evidence was unlikely to overturn the decision. Both of these involved highly value-laden issues -- the two most political issues encountered during this study, in fact. In both cases there were at least two opposing and strongly felt points of view, and in both cases the issue received a great deal of attention in the press. In one of the cases -- district 115 -- the controversy centered on a racial sensitivity training course that had been required of all school district staff. The staff was so incensed by the requirement that the issue eventually became an election campaign issue for school board members, and a new board which opposed this requirement was elected. This board changed the program's status from mandatory to voluntary, pending evidence which would indicate that the program was sufficiently effective to warrant its being mandated. The other episode, in district 35, involved a new back-to-basics elementary school which a group of parents had been pushing the district to initiate. The controversy in district 35 was primarily internal. The board felt pressured to comply with the request, while the administration felt that the presence of such a school would imply that there was something inherently wrong with the other elementary schools in the district. From the administration's point of view, the achievement test data for the district did not suggest a need for more emphasis on the basic skills. The board decided there should be a back-to-basics school, but that it should be evaluated. If the evidence indicated that it was harmful to students, it would be abandoned.

In both of these cases, the likelihood that the evaluation would yield the evidence needed to reverse the decision was slim. The history of educational evaluations is brief, but sufficient to indicate that differences among elementary education programs are difficult to discern, and that the discovery of substantial differences among them is almost unheard of. And racial attitudes that have developed over a lifetime are not likely to subside under the influence of a one-semester course. One would not, therefore, expect even a well-wrought study to provide either definitive evidence of negative effects of a back-to-basics school or definitive evidence of positive effects of a racial sensitivity training program.

Despite the way in which the issues were framed, the evaluators in both cases took their task seriously. In district 115, the board had required an external evaluation on the grounds that the issue was politically sensitive and there was a need to assure impartiality in the evaluation. The district evaluators worked closely with the contractors, and members of both staffs lost several weekends in their effort to complete their evaluation within the requisite time. In district 15, the study was done internally, and the evaluators devoted a great deal of attention to the design and conduct of their study, so that it would respond to the unique features of the back-to-basics school.

Neither decision was reversed when the studies were completed. The evaluation of the racial sensitivity training program was, at best, equivocal. Almost every finding highlighted in the summary was an "on the one hand/on the other hand" type of statement, and the school board took this evidence as supporting their position that the program did not deserve to be mandated. Apparently, the findings regarding district 15's back-to-basics school were not so equivocal, however, for that study was suppressed. Without access to it, we do not know what the findings were, nor do we know which faction in the controversy was responsible for its suppression. This was one of only two instances we found in which formal evidence was intentionally suppressed.[2]

The other three decisions which we label here as based on foregone conclusions involve decisions to adopt pilot programs in districts 57, 83 and 240. The stories are sufficiently similar that we will present only one, which involved a pilot Title I program initiated in district 240. Title I is a federally funded and regulated program designed to provide additional services to poor children who are behind academically.[3] School districts are required to demonstrate that their Title I funds have been used for additional services, rather than to pay for services students would have received anyway, but aside from accounting rules such as this, districts have considerable flexibility as to how they use these funds. They can purchase either new teachers or aides for existing teachers; they can serve children in their regular classrooms or they can pull them out; they can choose to emphasize reading, math, or language arts; and they can provide services to students in whatever grade levels they choose. Some districts change their Title I program quite often in an effort to determine the best way of using their Title I funds. District 240 did not have a history of changing its Title I program, but on this occasion it had come up with a new approach which several members believed would be an improvement. Before making the change, however, the district established a pilot version of the new program, and had its evaluation unit compare this version with the existing program. After a one-year trial, the pilot version was adopted throughout the district.

Described in such a succinct way, this sequence suggests that an old-fashioned rational decision had been made, and that the study was instrumental to that decision. But some important details were omitted above. One is that there were three reasons other than evaluation results for preferring the new program: the superintendent wanted it, it solved some complicated scheduling problems, and it created a clearer case that Title I funds were providing additional, rather than replacement, services. Furthermore, the decision to expand the pilot program was actually made before the evaluation data were available to indicate whether the pilot program was better or worse than the standard district Title I program. However, this sequence did not come about because participants did not care about the findings, but because they were convinced that the findings would be positive, and felt an urgent need to proceed with the plans for the new, expanded program. Finally, the new Title I program that was eventually adopted district-wide was not, in fact, the same as the pilot program. The pilot version provided children with either reading or math, depending upon which was the child's weakest area, but the new expanded program provided all children with an equal portion of reading and math instruction, regardless of their achievement levels in each area. Such a change could mean that the district-wide program would offer a weaker treatment than the pilot program had, but no evaluation of this new version was planned.

The remaining two episodes also involved pilot projects which were adopted or expanded prior to the release of the evidence. Adding these to those episodes described above, we have five examples in which systematic evidence was ostensibly instrumental to a major decision, but in all examples, the real decision was made prior to obtaining the relevant data. Yet we cannot say that the evidence served no purpose, for in all cases the decision makers appeared to have correctly estimated what the evidence would say, and in none of the cases in which we had access to the evidence did we find that it substantially differed from those expectations. Furthermore, it is not necessarily reasonable to assume that the one suppressed evaluation contained contradictory findings. The evaluators themselves were known by both camps to be opposed to the back-to-basics school, and their report could have been suppressed on justifiable grounds of bias.

In the absence of any examples in which late-arriving evidence contradicted the decision it is difficult to say that such findings would not have influenced participants to reverse their decision. But such examples may be hard to find, for not only would the evidence have to contradict the decision, it would have to be clear and compelling. Yet most "facts" can be interpreted in a variety of ways. And there are indications in these episodes that evaluators themselves either embraced the unified point of view that had already developed, or chose for some other reason to facilitate it. In the case of district 115's racial sensitivity training program the equivocation was so thorough and uniformly spread throughout the report as to suggest that equivocation was a conscious decision, rather than something that sprang from the data. And district 240's evaluators did manage to produce their findings regarding implementation of the pilot program in time for their clients to begin planning for district-wide adoption, even though they could not complete their analysis of the evidence on effectiveness. They, too, apparently assumed that the new version would be effective, and so concentrated on providing the information needed for planning the program's expansion. In one of the other cases involving pilot programs, the evaluators insisted in private interviews that their data were not consistent with the district's decision to adopt the pilot program. But their written report, as well as the oral presentations we observed, presented a careful balance between positive and negative findings, and they were liberally sprinkled with caveats regarding certain methodological aspects of the study, thus enabling their audiences to freely infer what they wanted. In this case, the formal decision was not actually made prior to the release of the evidence, but everyone involved, including the evaluators knew what the preferred decision was. If the evaluators believed the decision was wrong, they were anything but forceful in stating their case.

These episodes, then, are cases in which the decision was not only a foregone conclusion in the minds of the decision makers, but in some cases they were foregone in the minds of evaluators. The studies involved in these episodes were often presented in such a way that they could not help but legitimate the already-unified point of view. Indeed, the direction of influence could be the reverse of what we have been seeking: rather than evidence influencing a point of view, we find instead a point of view influencing the evidence.

Delayed Reactions

Just as decisions can precede the evidence, so can they develop some time after the relevant evidence has been presented. This section reviews the circumstances involved in three such cases. None of these delays were due to lengthy debate over the meaning of the evidence, however. Instead, the evidence lay dormant for some time and then was used. In all three cases, the stimulus that lead to the eventual use was a new political struggle, and in all three cases, the old, previously ignored evidence was drawn on in connection with new political issues.

One of these occurred in district 50, and involved a Follow Through Program. Like Title I, Follow Through is a federally funded program designed to serve poor children. Beyond that common feature, however, the two programs are quite different. Follow Through tends to be a full-day program, rather than offering an hour or so of extra help, and it provides a number of non-instructional services to children -- medical examinations, inoculations, dental care, and so on. For a variety of reasons, the national budget for Follow Through has always been very small relative to Title I, and it had been getting smaller and smaller in the years immediately preceding this study. The effect of these national changes in district 50 was to reduce its program from six schools to four classrooms.

Like most federal programs, Follow Through required school districts to annually evaluate their programs. District 50 had routinely conducted annual evaluations of its Follow Through program, and these evaluations had been routinely negative and routinely ignored. The program director believed that the data were not relevant to the program, since the program was designed to increase children's abilities on attributes not measured by the district's standardized test. Her supervisor, a member of the superintendent's cabinet, believed the program was valuable for reasons unrelated to test scores: parents liked it, and it provided medical and dental services that poor children really needed.

The year before we visited district 50, the district desegregated, and in the process the four remaining Follow Through classrooms were moved to new school buildings. New principals thus came into contact with the program and one of them didn't like it. She wanted to modify these classrooms, but in order to do so she needed authority to supervise the Follow Through teachers. She entered into a power struggle with the program director, and the struggle eventually tumbled into the superintendent's cabinet. The superintendent began asking questions and everyone was reminded of Follow Through's evaluation history. Two members of the cabinet even volunteered this episode to us as an example in which a school district had not responded to its own evidence.

Follow Through was suddenly a hot topic, and what to do with it a big question. The cabinet member overseeing the program asked the evaluators to compile a major review of all the evidence regarding the program, and to organize it around a set of predetermined questions. The cabinet as a whole decided to have a blue-ribbon panel review the program and make recommendations. Most of the participants were reluctant to abandon the program because it would mean a loss of federal funds. Eventually, they decided to keep the program but to change its curriculum, and to appoint an independent monitor to see that the changes they wanted were implemented. Around the time that this decision was made, we completed our field work in district 50. However, three months later we learned that the district had decided to discontinue its Follow Through Program.

There are several plausible reasons why this decision could have been made. First, the program was philosophically different from the other educational activities in the district. It was an open-classroom program designed to foster self-reliance and independent problem solving, while the remainder of the district had become increasingly interested in a uniform structured curriculum which guaranteed that every child would attain the same basic skills. Second, the director was not inclined to cooperate with the changes the cabinet had been trying to impose on the program. Third, the evaluation data were negative, and had been consistently so for several years. Fourth, the program was responsible for a bothersome power struggle. And finally, it was a small program and was getting smaller every year, so that its size hardly justified the headaches now associated with it.

How much of the decision was based on the evidence is hard to say; however, the evidence had clearly been overlooked for several years, and was clearly not the triggering stimulus for the change. In fact, its contribution came about only because a new configuration of people and ideas encouraged members of the district to reconsider the program and to reinterpret the evidence about it. In earlier years, the program had been viewed as having several non-instructional advantages, as bringing new ideas and new money into the district, and as offering children a unique learning environment whose effects could not be measured by ordinary test data. Under the new configuration of events and people, the program was viewed as inconsistent with district-wide practices, too bothersome to justify its meagre financial contribution to the district, and as lacking any special academic advantage. In both periods, the evidence was part of the program's image, but the meaning of the evidence changed from one image to the next.

Of the two remaining examples in which old evaluations became part of new issues, one will be reviewed only briefly, and the other will not be reviewed at all. The latter involves a court case, and will not be reviewed because a description of events may reveal the identity of the school district. However, the court-case episode differs from the others in this category in one way that needs to be mentioned: the old evaluations indicated that the program had positive effects, rather than negative. We mention this fact here to indicate that positive program effects can also take on new meanings and hence be put to new uses. In this case, the court ultimately ordered the expansion of a program that the district had been operating complacently for many years.

The third example in this category involves negative evaluation data and a bilingual program in district 83. The particular events involved in this episode are different from those in the Follow Through story, but their pattern is similar in that the program continued in a relatively stable status until there was a change in the pattern of relationships among members of the district. In this case, the change came with a new school board and the program was not dropped but was instead revised considerably. The episode also illustrates the variety of reasons that can be given for a decision. Participants in this decision listed four reasons for revising the bilingual program. First, the district was engaged in a variety of efforts to unify the curriculum, and this was merely one more effort along that line. Second, it was time to move the program from its early developmental stage of social protest to another developmental stage, called "educational pedagogy." Third, the district hoped to obtain funds from the state, and these revisions would make it eligible for those funds. Finally, the program was poorly managed and poorly staffed, and it simply "needed work." These, then, were the justifications for revising the program. They are not presented here in their relative order of importance, for no preference was indicated by interviewees. In addition to these four reasons, there were two others offered by members of the district who did not approve of the changes. One was that there was a power struggle between the superintendent and the school board, and it happened to settle on the bilingual program. The other was that the decision was based on negative evaluation data, and that the evaluation had been a hatchet job. We were unable to obtain either of the two relevant evaluation studies -- the earlier one was out of print and the later one had been done by a contractor whom we did not have a chance to contact. The two studies were said to have similar negative findings.

Though we have only three examples of delayed reactions to evidence, the examples we do have share certain features. First, the delay was not the result of an unusually lengthy process of negotiation or other participatory processes needed to unify points of view. Instead, members were already unified when the evidence was released and they agreed that the evidence did not indicate a need for change. Second, the evaluations themselves were not sufficiently compelling to stimulate new perceptions of the programs. Instead, they lay dormant until power shifts stimulated people to re-think the programs. Third, in all three cases, the shift encouraged participants to view several aspects of the program -- not just its effectiveness, but, for instance, its connection to other district programs and practices -- in a new way. Finally, these new perceptions also caused the evaluation evidence to be interpreted differently than it had been before, so that it suddenly had implications for action which it had not had before.

Provoked Reactions

Our last category includes six episodes in which the evidence appears to have been responded to soon after it became available This is the temporal sequence which we would most expect to indicate that the evidence had an independent substantive effect on decisions. The distribution of examples in this category is somewhat unusual, since four of the six episodes involve evidence of changes in demographic data, and these four cases all came from a single school district. We will therefore review these four as a group, following review of the other two episodes.

The first episode, which occurred in district 115, is closest of all the episodes to an example of old-fashioned rationality, in that the evidence was instrumental to a decision and there appeared to be no other motives for the decision. Members of this school district had had, for several years, a concern that the high school students were cutting classes too much. So, two years before our visit, the school board initiated a new policy such that if a student accumulated 10 or more unexcused absences in a class, he or she would lose credit for the course. In order to accommodate the students' right to due process, the policy also stipulated a series of warning procedures that were to be implemented by teachers -- an informal meeting with the student after one absence, a notice to the principal after three, a letter to the parent after six, and so forth.

As time went on, however, the high school principals felt dissatisfied with the policy, and asked that it be evaluated. Not only did the policy require a considerable amount of paperwork, but it wasn't clear that it was deterring students from cutting classes. Some of the principals also hoped to learn something about the sources of absenteeism -- what types of students tended to be absent, and so forth.

The evaluation unit conducted a relatively thorough study of class-cutting, though the results, like the results of almost every study conducted in this district, were designed to inform the board, not the principals. The report indicated the distribution of credit losses among students as a function of sex, race, grade level, and school building. In addition, it indicated:

- a high correlation between how aggressively the policy was implemented and the school's attendance -- a relationship interpreted to be causal;
- an estimate from principals that they needed an additional 1/2 person to maintain attendance and procedural records;
- the finding that principals thought students interpreted the policy as permitting up to nine free class cuts.

The last of these findings was particularly surprising to the board members. Apparently it had not occurred to them that students might perceive the policy as permitting, rather than restricting, class-cutting. In response, they revised the policy, reducing the number of permissible cuts from ten to five. In view of the facts that (a) this study was conducted specifically to inform a decision about this policy, (b) a decision was indeed made, (c) the study was the primary basis on which that decision rested, and (d) there appeared to be no political factions or value-laden conflicts involved; one could say that this decision was scientifically rational.

However, much of the study's findings were not attended to. For instance, the fact that the policy was effective when implemented could have been construed to mean that no change in the policy was needed, but rather that emphasis on implementation was needed. And the implementation problems associated with the policy's procedural requirements were not addressed. In fact, the new policy exacerbated those problems, since the required sequence of steps was not abandoned but rather hastened, so that all the steps could be carried out within the space of five class cuts rather than 10. The board did not, then, consider either the effectiveness of the policy or its implementability, but instead reacted to the hearsay finding that they themselves were perceived as lenient. Their decision showed students that they were not lenient.

Our second example comes from district 25. District 25 has a relatively large population of students with Hispanic backgrounds. Some have been in this country long enough to have developed reasonable facility with the language, but others are new and have only minimal, if any, understanding of English. Many are migrants, and do not spend the entire school year in district 25. Because of this population, the district offers a number of special bilingual programs--programs for migrants and non-migrants, for Spanish-dominant and English-dominant, and for younger and older children. Among these programs were two, one in oral English and one in written English, which were offered to students in grades three-through-six and one-through-three, respectively. The oral English program was locally developed, and the district was proud of it. To evaluate it, the district used the Language Assessment Survey (LAS), an instrument it learned about when the U. S. Office for Civil Rights required that it be used to assign students to language programs appropriate to their own language fluency. Several members of the district complained about the LAS on the grounds that it was administered too often and that students could learn the correct responses to questions without necessarily understanding the questions themselves. Students in the written English language program were evaluated with a norm-referenced, standardized test, one used by the district for all of its routine testing needs.

Each program was evaluated against an intended gain score. Students were tested on their respective tests at the beginning and at the end of the school year, and the findings indicated that, across grade levels, the percent of students in the written English program who met the criterion ranged from 16 to 32 percent, and the percent in the oral English program who met criterion ranged from 41 to 100 percent. The district took these data to mean not only that the oral English program was superior to the written English program, but also that oral English should precede written English in the grades. This latter conclusion apparently was stimulated partly by the data and partly by the observation that teachers of earlier grades were more skilled in teaching oral language than were teachers in the later grades. District administrators then initiated plans to expand the oral English program down into the earlier grades.

Even though some members of the district had complained about the weaknesses of the LAS, others praised it as enabling them to see the superiority of the oral English program to their written English program. No one mentioned any of the issues related to the fact that the two programs were evaluated against different outcomes: i.e., that one outcome might be more difficult to achieve than the other, that one test might be less sensitive to their curriculum than the other, or that scores from one test might have been more susceptible to practice effects than those of the other. Instead, the evidence was assumed to reveal comparative benefits of the programs and program redesign followed from that interpretation. Furthermore, even though several interviewees attributed their programmatic changes to these evaluation data, several also mentioned that they were particularly proud of the oral English program because they had created it themselves. To the extent that participants wanted to believe the program was successful, this could be an example of a foregone conclusion rather than a reaction provoked by the evidence. On the other hand, it is not clear that anyone without psychometric training should be expected to interpret such widely different percentages of successful students as indicating anything other than differentially successful programs. To the extent that their interpretations were genuine, the episode belongs with others in which the evidence is directly responded to.

Let us turn now to the remaining four episodes, all of which came from district 7. The population of district 7's community has been changing gradually but perceptibly for the past 15 years, from one of predominantly educated, white-collar workers to one dominated by senior citizens and blue-collar workers. The change has had two noticeable effects on district 7's student body: it is much smaller than it was fifteen years ago, and fewer of those students who are there plan to attend college.

District 7 is too small to maintain an evaluation office. However, it does maintain test data and enrollment data, and the high school guidance office conducts occasional surveys of the community, usually in response to the accreditation requirements. But although the district has far fewer data at its disposal than do many of the larger districts in this study, it appears to take full advantage of the data it has. Three of district 7's four episodes that fall into this category have to do with such descriptive statistics as enrollment trends or trends in the composition of the student body. These data stimulated the district to create new programs for adults and preschoolers, for instance, in the hope that by doing so it would not need to close any of its buildings as student enrollment decreased. And the data stimulated the district to increase its commitment to vocational education. Though the data themselves were primarily descriptive, they were taken to have some rather dramatic implications, and over the past 15 years, several major changes have occurred in district 7. However, because these changes occurred prior to our visits, the precise relationship between the evidence and the decisions is difficult to ascertain. Therefore, only the most recent of the episodes that occurred in district 7 will be described. Though less is known about the others, they appear to be similar to this in those respects that are relevant to our interests here.

This episode, like the others from district 7, involved the derivation of new courses of action from data that were essentially indicators. The issue in this case was whether or not to offer algebra to eighth grade students. Algebra had traditionally been available only to students in ninth or higher grades, and there were two lines of reasoning for why it might be moved down a grade. One was that, since the district had begun offering preschool, some of the curriculum was beginning to slide downward. As the preschool adopted kindergarten materials, the kindergarten picked up first grade things and so on. Therefore, by the time students reached eighth grade, they would be ready for ninth-grade coursework. The second had to do with the fact that high-school students wanted more flexibility in their schedules, and their flexibility would be increased if they could begin their math sequence a year earlier. The second line was put forward primarily in relationship to gifted students.

The idea of making algebra available in the eighth grade apparently first came from parents. The superintendent had liked it, however, and had asked one of his staff members to gather some relevant data. This staff member conducted a telephone survey of 21 neighboring school districts and found that 17 of them offered algebra in the eighth grade. She also provided him with data on the number of students in sixth and seventh grades whose standardized test scores were at or above the 90th percentile, and some other statistics regarding the distribution of intelligence quotients, course grades, and the like, that could serve to indicate how many students might be permitted to enroll in such a course.

Like most school districts, district 7 employs a complex committee system for many of its decisions -- recall that it is district 7's committee structure that is displayed in Figure 1. The algebra decision was formally the province of the math curriculum planning committee, so the process began when the superintendent took the idea and the evidence to this committee. Meantime, an assistant superintendent took the idea to several other committees: the parent advisory committee, the student advisory committee, and the special education advisory committee. But while these two were generating a broad base of support for the idea, members of the math committee were talking it over with their colleagues in the schools. They found resistance to the idea, and they, as a committee, rejected it. After some negotiation, however, it was agreed that the course would be offered two years hence. There were two rationales for the delay. First, students who had had the preschool program were not yet into the eighth grade (though they also would not be there in two years, according to our estimates of when the preschool program began), and second, the delay would give teachers some time to prepare for the change.

Though the sequence indicates some features of district 7's decision-making practices, such as the use of ad-hoc analyses of extant descriptive statistics and the procedural use of committees, it also indicates, like many of the episodes described earlier, the role that participatory processes can play in the outcomes. In this case, participants mentioned in private interviews that the main reason the committee balked at the suggestion of eighth-grade algebra was that it was too clearly the superintendent's idea, rather than its own. Under normal procedures, for instance, the committee, rather than the superintendent, would gather the appropriate evidence. When the superintendent, in his zeal, took the liberty of doing the advance exploration on his own, he was usurping the committee's role. And, since the superintendent had apparently agreed at one time never to take things to the board without the concurrence of these curriculum action committees, he was unable to move without this group's recommendation. Several interviewees referred to the incident as one designed to establish the committee's power and to teach the superintendent a lesson. Yet despite this apparent muscle flexing, the superintendent did get the decision he wanted, and given the nature of his supporting evidence, it is not likely that the committee was compelled by evidence.

The six episodes gathered together under the category of provoked reactions all involve relatively immediate responses to evidence, but they also all involved rather creative interpretations of the evidence. In district 25, two independent evaluations, using different outcome measures to evaluate different programs serving different age groups of children were interpreted as indicating the comparative advantages of the programs, and furthermore to indicate that one of the programs should be offered to children of a different age than that of the group on whom the program was evaluated. In district 7, descriptive data regarding student body characteristics were used to infer the need for a variety of programmatic changes. And in district 115, when the school board decided to reduce the number of unexcused absences leading to denial of course credit from ten to five, it was responding primarily to hearsay to the effect that students interpreted the former policy as lenient, rather than to the evidence regarding the actual effects of the former policy. The fact that responses followed quickly on the release of the evidence, therefore, while indicating that the evidence may have influenced the unification process, does not indicate that it had an independent substantive effect on decisions. The evidence was not compelling because of its scientific merits or its indisputability, but rather because the already unified inclinations of the participants led them to interpret the evidence as compelling.

SUMMARY AND CONCLUSIONS

Our interest in the episodes presented here stems from an interest in understanding the possibilities inherent in a compromise between participatory and scientifically rational decision-making, and in particular with answering the question of whether and how evidence can be used in decision-making contexts that are inherently participatory. The ideal role for evidence in such settings is difficult to define for two reasons. First, the decisions that are made in these contexts are rarely as straightforward as the traditional conception of scientific rationality would have them be. They may involve several intertwined issues and they may be resolved by political, social, legal, or organizational processes, as well as by reliance on evidence. Second, once we acknowledge the legitimacy of these several participatory processes, and concede that participants may have differing values and interests and that they may interpret evidence accordingly, then the evidence no longer is indisputable and the ultimate decision no longer needs to have any particular substantive relationship to it. A review of real examples, then, should shed some light on whether or how evidence could influence either the processes or the outcomes of participatory decisions.

Each of the categories of decision sequences that we reviewed here has provided some insight into this issue. The first consisted of those cases in which decisions preceded the evidence. These were situations in which the political or social processes produced a unified point of view so forceful that the evidence could not have served a purpose other than to reinforce it. If the evidence had shown clearly and compellingly that the decision was wrong, we might have thought of these as social movements verging on mass hysteria, but in these cases either the decision makers were right or nearly right in their estimates of what the evidence would say, or they were able to convince the evaluators of their views so that the evidence actually put forward by evaluators confirmed their views for them. These decisions did not merely precede the evidence, then, they anticipated it. The second group of episodes contained studies which had delayed effects. These delayed responses turned out not to be the result of lengthy participatory processes which followed the release of the evidence, but instead were cases in which, although the evidence remained constant over time, its perceived message changed. In all three cases, the new perceptions developed when new configurations of decision-making participants stimulated a host of new interconnected ideas. The eventual decisions appeared to be consistent with the evidence, but did not derive from it. Instead, they derived from the new configurations of people and ideas. The third category of episodes contained those situations in which participants responded to the evidence soon after it was released. These were the decisions on which we most expected to see an independent effect of evidence. And indeed, all six episodes arose with the introduction of the evidence and concluded with one or more programmatic changes which appeared to be consistent with the evidence. Yet in all of these cases, the evidence itself was either sufficiently ambiguous or sufficiently complex that it could be interpreted to indicate that no changes were needed or that changes other than those that were made were needed. Thus, although the evidence may have stimulated the decisions, the substantive concurrence between evidence and decision was due more to the interpretive predilections of decision makers than to any inherent truth contained in the evidence itself. Apparently, the political, social, or organizational processes that preceded the release of the evidence had been sufficiently successful at unifying participants that they were all inclined to interpret ambiguous evidence in the same way.

The three sequences reviewed here cover the range of possible temporal relationships between evidence and decision, but there is one other fate that evidence itself may have: it may be ignored so completely that it is never mentioned by anyone, either in meetings or in interviews with visitors. It is possible that more studies meet this fate than any of the temporal fates described here, and in fact, during the course of our study, we came upon copies of a great many evaluation reports which were never mentioned by anyone, except perhaps by evaluators who provided them as examples of their work. The frequency with which this fate occurs, relative to others mentioned above, however, is difficult to determine, for the fact that no overt decision is made, nor any overt reference made to a study, cannot be taken to mean that participants are not aware of the study or its findings. And if they were aware of the findings, the chances of use are considerably increased, even though the use itself may not be soon or visible. If the findings suggest that current practices are effective, for instance, or even acceptable, decision makers may simply acknowledge these findings and decide not to open the program up to scrutiny. If, on the other hand, the evidence challenges current practices, decision makers may still acknowledge the findings, but do so in a way that permits them to infer no need for change, yet which also permits them to draw on the evidence at any time in the future when other events may suggest a need to review and modify the program. If we recall our three examples of delayed reactions, we find that these studies were, in their dormant stages, interpreted as either consistent with current practices, in the case of the positive findings about a program, or as invalid or irrelevant, in the case of those findings that indicated that the program was not very effective. Later they were interpreted to mean that changes were needed.

The dormant and non-dormant stages in the histories of all these studies are really very similar. In all cases, the studies' interpretations, and consequently their uses, were determined by the prevailing views of decision makers, and these in turn were molded by political, social, or organizational participatory processes. Even when the evidence appeared to provoke immediate reactions, the decision makers were not moved by clear and compelling evidence of a need for change. Instead, they rendered the evidence meaningful to their views, and only thereby did they perceive a need for change.

Advocates of a compromise model of rationality seek a balance between the need for participatory decision making, on the one hand, and the need for scientific rationality on the other. This review was intended to illuminate the various ways in which these two forms of decision making could be merged. The problem posed at the outset of the paper was this: to the extent that evidence has no independent influence, and is instead merely a tool that is used either well or poorly by participating debaters, negotiators, or bargainers, then whatever resolution develops will not be determined by the substance of the evidence but rather by the distribution of political and manipulative skills among participants. In none of these 14 highly diverse examples of decisions did we find a situation in which evidence appeared to have an independent influence. Instead, the participatory processes created a unified point of view that was so compelling that it imbued the evidence -- as well as a variety of other circumstances -- with a meaning that was consistent with itself. It seems reasonable to ask, given these findings, whether evidence can provide an independent contribution to decision making. Some authors -- most notably Lindblom and Cohen (1979) -- have argued that if it does, its contribution will necessarily be marginal. And the episodes presented here support that argument. They suggest that the consensus that evolves through participatory processes is by far the more powerful influence on decisions. In fact, it appears that prevailing ideas drive the evidence, rather than the evidence driving the ideas. When people say they have used evidence, what they really mean is that they have rendered it meaningful, by connecting it to a prevailing and usually very powerful point of view. Having done so, they can claim the evidence is relevant, timely, and compelling.

Such a conclusion may be jarring to organizational observers who, though aware that many studies lie dormant and that many others are interpreted creatively, nonetheless are also aware of cases in which participants claim to have, and appear in fact to have, made a scientifically rational decision on the basis of the evidence. Indeed, critics to this conclusion could point out that none of the decisions reviewed here appeared to clearly contradict the evidence associated with them. But even if evidence had no independent effect there would still occur, within a population of decisions, some randomly distributed examples of substantive correspondence between evidence and decision. It is entirely possible that a study such as this retrieves only those occasional random artifacts of decisions which are entirely controlled by ideas that evolve from participatory processes. In these cases, participants may say that their decisions were based on the evidence, but what they really mean is that (a) the prevailing point of view happened to evolve in such a way that (b) a particular interpretation of the evidence happened to be both palatable and possible, thus enabling (c) a particular programmatic decision to appear to be rational rather than serendipitous.

NOTES

[1] District code numbers indicate the approximate size of the district in thousands of students. For instance, district 7 serves about 7,000 students and district 83 serves approximately 83,000 students. Actual enrollments randomly deviate from the code-implied enrollments by ±15 percent.

[2] The other instance involved enrollment data indicating overrepresentation of blacks in special education classes in another district.

[3] In 1980, Congress repealed several earlier education provisions and established new ones. As a result, what was Title I of the Elementary and Secondary Education Act is now Chapter 1 of the Education Consolidation and Improvement Act. The episode described here occurred prior to this change.

REFERENCES

Allison, G. T. Essence of Decision: Explaining the Cuban Missile Crisis. Boston: Little, Brown, 1971.

Cohen, D. K. and Weiss, J. Social Science and Social Policy: Schools and Race. The Educational Forum, 1977, May, 383-413.

Cook, T. D., Levinson-Rose, J. and Pollard, W. E. The Misutilization of Evaluation Research: Some Pitfalls of Definition. Knowledge: Creation, Diffusion, Utilization, 1980, 1, 477-499.

Deising, P. Reason in Society: Five Types of Decisions and their Social Conditions. Urbana, IL: The University of Illinois Press, 1962.

Ezrahi, Y. Utopian and Pragmatic Rationalism: The Political Context of Scientific Advice. Minerva: A Review of Science, Learning and Policy, 1980, 18(1), 111-131.

Lindblom, C. E. The Intelligence of Democracy: Decision Making through Mutual Adjustment. New York: The Free Press, 1965.

Lindblom, C. E. and Cohen, D. K. Usable Knowledge: Social Science and Social Problem Solving. New Haven: Yale University Press, 1979.

Lyon, C. D., Doscher, L., McGranahan, P. and Williams, R. Evaluation and School Districts. Los Angeles, CA: The Center for the Study of Evaluation, University of California Graduate SChool of Education,1978.

March, J. G. and Olsen, J. P. Ambiguity and Choice in Organizations. Oslo, Norway: Universitetsforlaget, 1979.

Simon, H. A. Administrative Behavior. New York: The Free Press, 1957.

EVIDENCE AND MANAGEMENT

If the past decade of education were to be characterized by a concern for meeting the unique needs of diverse students, the current decade could equally well be characterized by a preoccupation with uniformity of services and with efficiency. School districts which decentralized their curriculum and decision making in the sixties and seventies so that each school could reflect its own neighborhood's interests are now centralizing their decision making and establishing uniform curricula across the schools. These changes have come about in part because parents and communities have not been impressed with their school's performance over the past few years and in part because school districts have less money and fewer students than they had a decade ago. The public has demanded fewer frills and better student performance in the basic skills and the budget has demanded that educational services be streamlined.

One response to these demands has been to increase the emphasis on test scores. Because instruction occurs in a multitude of separate schools and classrooms, educational managers often base their oversight on other forms of evidence about what is occurring in the schools. Tests not only facilitate oversight, they also facilitate a number of other management strategies. Tests can be used to define the curriculum and to communicate to teachers what they are expected to teach; they can form the basis of a system of management by objectives; and they can be used to measure the progress of teachers and schools and to identify those who need in-service training or other forms of assistance. They offer a powerful tool with which administrators can not only improve their knowledge of what is occurring in their schools but can also influence practices within the schools. They can make instructional decisions at all levels of the district more rational.

But tests can only increase the rationality of education if they can also improve the rationality of decisions inside the schools. Teachers and principals are expected to attend to their students' performance on the test and to develop new practices that will improve their students' test performance. Managers who use tests for oversight generally hold two interrelated sets of assumptions regarding how their management system will or does influence school and classroom decision making, and both sets of assumptions also have to do with the nature and value of rational decision making. The first set of assumptions has to do with the extent to which school principals and teachers are motivated to be rational. This set includes three assumptions. The first is that if evidence is used at all, it will be used in a scientifically rational way. Therefore, requiring the use of evidence will increase the extent to which school and classroom decisions are rational. The second assumption is that, because the system formalizes the responsibilities of each individual in the educational process and makes each more accountable, it therefore motivates each to choose the most effective means for improving his or her own performance. The third assumption in this set is that because the evidence defines the goals of education and makes them more visable, it enables staff to focus their efforts more precisely. Taken together, this set of assumptions leads to the belief that management systems that rely on tests will promote scientifically rational school and classroom decisions. The second set of assumptions has to do with the extent to which school principals and teachers can make scientifically rational decisions. This set includes two assumptions. The first is that the meaning of the test data is self-evident -- that principals and teachers can readily decipher the patterns in the data and "see" what needs to be done. The second assumption is that principals and teachers have the technical knowledge and skills needed to provide their students with whatever knowledge they still need to learn. Once teachers know what their students know, don't know, or need to know, they will be able to change their practices accordingly and fill in the gaps. Taken together, these two assumptions lead to the belief that a management system that emphasizes improvement in test scores can be used by school principals and teachers to make more rational decisions about their practices.

Administrators who use tests in this way have shown little interest in evaluating the effectiveness of their management systems. Because the systems themselves are considered to be methods for improving the rationality of instructional decision making, and as encorporating evaluation tools within them, they are not viewed as in need of evaluation themselves. But because they are based on such a complicated array of assumptions, their actual effect needs to be compared to their assumed effect. For it is possible that if even one assumption is false, the entire system may not work as intended.

This paper provides an initial and highly tentative evaluation of the family of management strategies that rely to varying degrees on test data. The data on which this evaluation is based come from a study of school district uses of evaluation and test data in which 16 school districts participated. These districts varied on a number of demographic characteristics such as size and ethnic composition of their student bodies, their wealth and their geographic locations. Data gathered from these districts was entirely qualitative, coming from interviews and from observations of group meetings. The management strategies discovered in these districts that relied upon test data can be loosely grouped at four points along a continuum according to the amount of emphasis they placed on tests (see Figure 1). At one end of the continuum are strategies we call "consultation" -- strategies in which administrators, usually principals, meet periodically with teachers to review test data and discuss ways of improving children's performance. Add more stress and we have "instigation" -- a set of strategies designed to motivate teachers by increasing competition among them. Finally, on the far right, are evaluation strategies, wherein teachers' annual evaluations are based to some extent on their use of tests. As we move from left to right along this continuum, tests should become increasingly more important for teachers, or at least more difficult to ignore. If the assumptions underlying these systems are correct, we should expect instructional decisions to become more rational as we move from right to left along this continuum.

These categories should not, of course, be taken as entirely discrete -- the real distinctions between these different methods of applying stress are quite blurred. For instance, one might expect that advice would be less anxiety-provoking than evaluation, but a particular consultation could be extremely stressful with the right combination of personalities involved, and with the right setting and circumstances. Furthermore, our evidence of efforts devoted to these various efforts suggests that they are cumulative. We could often find examples of consultation in isolation, but examples of efforts further along the continuum usually were accompanied by examples of less stressful efforts as well. Many of these difficulties of definition, we hope, will dissolve as we explore in more detail our family of strategies. After reviewing these four strategies for increasing attention to tests, and their effects on teachers, we will discuss some issues related to the overall effectiveness of management strategies like these.

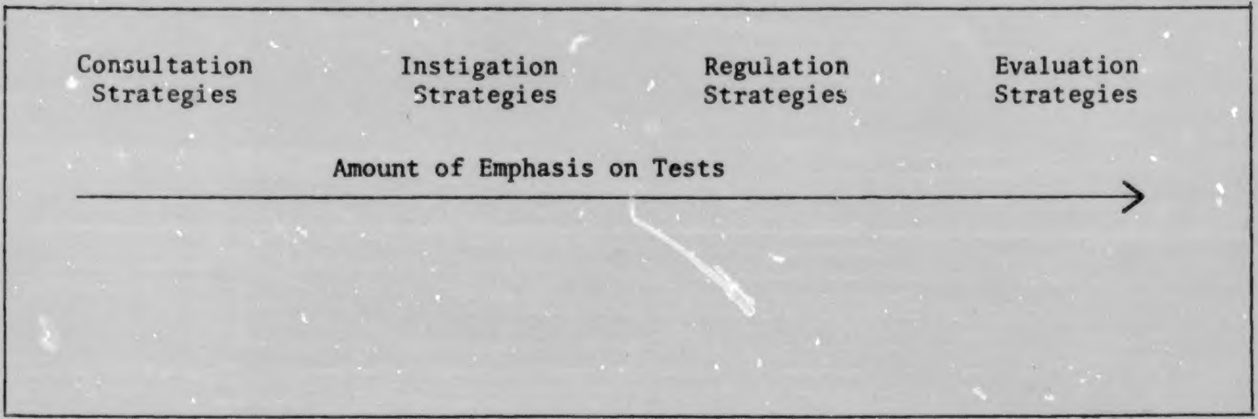
Consultation Strategies
Instigation Strategies
Regulation Strategies
Evaluation Strategies
Amount of Emphasis on Tests

Figure 1: The family of organizational strategies designed to increase teacher use of tests.

CONSULTATION

We use the term consultation to refer to methods of helping teachers to analyze test score patterns in order to modify instruction. A consultation usually involves a principal and either one teacher or a small group of teachers, with the principal serving as the counselor. He or she gives advice regarding the interpretation of test results and guides teachers in making appropriate decisions based upon those test results. A faculty meeting where all the teachers in the school are present and where the principal presents the building's annual test scores would not belong in this group unless the principal uses this opportunity to promote buildingwide soul-searching regarding curriculum or pedagogy. The following comments made by one principal illustrate what we would call a consultation to promote the use of test results by teachers. The principal explained to us how he reviewed test scores with his teachers:

> I look over the test results for any kind of anomaly. Then I will discuss with that teacher the anomaly, whether it is a good score for an otherwise poor student, or a poor score for an otherwise good student. I relate to the teacher my thoughts on the test results. I also talk to the teachers about the general average of the class. Not long ago I shared concerns with my first grade teachers. We came up with two classes; one with an average [IQ] of 112, and then we have another class with an average [IQ] of 104, and both of those classes scored in the low 20s [results on a standarized achievement test]. [District 7]

The mood of such consultations tends to be one of collegiality -- a group of concerned educators reviewing evidence regarding the effectiveness of their practices. And, for the most part, teachers seemed to feel they profited from these exchanges. Witness this teacher reaction:

> This year the principal reviewed the [test scores] with the second and third grade teachers to determine weaknesses of the program. We broke it out question by question. . . . This forced me to change my curriculum. I was teaching some of the tested skills after the tests were given. In other cases, such as the differences between fact and fiction, I simply was not teaching this point. [District 7]

Though the two comments cited above came from a fairly small school district, the process can occur in larger districts as well. The quote below is from one of the five largest districts in the country:

> [The principal] takes this [looseleaf notebook with math and reading scores] monthly . . . and he asks you how the kids are doing, and he asks about each of the kid's scores in the book. He collects the books monthly, and he always sends a reply. He doesn't harass you; it's a help actually. To me, it helps structure things; it shows that he is interested in the kids. That's the way it should be. Sometimes it's a pain when he says to cover six chapters this term. He's a unique individual. If I were an administrator, I would do the same thing. He's not in the classroom observing, mainly he keeps tabs on you through reviewing the book. [District 220]

These examples are illustrative of the mood and the felt benefits of consultations. Relative to the methods we will discuss later, these consultations place less emphasis on tests, and (as far as we could tell) they did not foster much anxiety among teachers. The principals who engage in these activities were apparently able to effectively guide and advise teachers in their use of test scores without using intense pressure to improve performance.

But the consultations we observed were, for the most part, done on the initiative of individual principals, not as part of coordinated districtwide activities. And, since the success of consultations depends upon the knowledge, skills, and personality of the principal, it may not be possible for such practices to become district policy. Before a principal can effectively guide teachers in their use of test scores, he or she must possess both an inclination and an ability to analyze and interpret test results, and many principals do not. This is not to say that most principals are inadequate, but rather that their skills may be in other areas. Further, even if a heavy investment were made in training principals in the needed abilities, it is not clear that their inclinations would change.

One reason their inclinations may not change is that these activities take time, and the time spent on these activities could otherwise be spent on other activities. The principal who collects monthly records, when asked about the time involved, said:

> I don't go around the building and look at bulletin boards or the condition of classrooms and complain that they should be picked up. When I go over these records, I don't do anything else. I work from 8:00 to 5:00 and don't leave at 2:45 when the school day ends like a lot of other people. I let everything else go until I finish going over the results and responding to every teacher in writing.

Furthermore, it is not clear that consultation increases the rationality of instructional decision making. It is more of a brainstorming activity during which alternative perspectives on the data can be put forward and considered. Whether teachers are motivated to actually change their practices is not clear. Though some described benefits of these consultations and described new insights they had into their own practices, few described specific changes they had actually made in their practices.

INSTIGATION

The term instigation refers to administrative efforts designed to motivate teachers not by guidance but by increasing competition among teachers. These efforts can be quite subtle, or they can be overt. A subtle form of competition was fostered by one principal simply by making class differences in test performance available to teachers. He said, "I don't lead staff meetings by making comparisions, but I don't discourage teachers from making the comparisions themselves. The information is available" [District 7]. A more overt form occurred in one large district in which teachers that had shown the largest increases in class performance over the previous year's performance were designated as "master teachers," and the district superintendent personally visited the master teachers and awarded them commmendations. In this way the district was encouraging teachers not only to compete with each other, but also to compete with themselves to improve over last year's test scores.

One principal we met was quite direct in encouraging competition among teachers: she calculated t-tests on the pre- and post-test scores of each teacher's students, and noted which teachers had made significant gains. She then produced a table with all teachers' results and distributed it to all teachers in the school. As it turned out, this method didn't have much effect on teachers. None of them understood the data.

One way districts can foster competition among teachers is through public release of test results. Though most districts release test results to the newspapers and other media, some districts have intentionally released test scores in such a way as to encourage comparisons among buildings and even among teachers within the same building. One teacher recalled what had happened the first time her district released test results in such a way:

> I remember the first year they had state-mandated reading tests. The teachers felt the pressure to succeed. . . . You see, the scores were printed in the paper. . . . This prompted competition between schools. The tests shouldn't be competitive. It causes cheating and competition. [District 27]

Several of the districts we visited had had some experience with releasing building comparisions in the past, but most had decided this was not helpful, for one of two reasons. The first had to do with the amount of pressure exerted on staff. For instance, an associate superintendent said, "Competition within the district between schools is healthy, but if [comparative test data] were published it would be too much competition for the district to bear" [District 35]. The second reason had to do with the adverse effects on the administration itself. Citizens often drew unwarranted conclusions from these comparisons, and translated them into almost impossible demands on administrators. Although district administrators perceived these demands as unreasonable, their experiences did not sway them from feeling that they should make similar demands on teachers, using essentially the same kind of comparative evidence.

Competition may be part of the American way, but it is not without its disadvantages in education. Not only does it decrease the extent to which teachers will work cooperatively together, but it places them in the middle of a zero-sum game: for there to be winners, there have to be losers. As one secondary school principal said, "Well, naturally if you are high in the rankings the effects are good, but I've been in schools on the other end of the spectrum and its effects are very depressing on you and the kids" [District 220].

These motivational strategies stress tests more than consultations do, simply by virtue of the evaluative judgments that are implied in these comparisons. But as far as we could tell, the added stress did not affect teacher behavior. We hypothesize that the failure of these strategies to induce behavioral changes based on test scores was due to the fact that, with the exception of those principals who consulted with teachers on their own, this strategy of stress inducement was not accompanied by any guidance as to what to do about low test performance. The message was only to raise scores; how to raise them was not addressed.

REGULATION

By regulation we mean the creation and implementation of a formal system which requires teachers to respond to test scores in some prescribed fashion. Several districts in our sample regulated teachers' use of test results so that responding to test scores became something teachers could not avoid. Two districts had procedures wherein teachers at the same grade level were required to meet and discuss annual test results and to prepare a written report analyzing the results and outlining what action they intended to take based upon the test results. Three others had adopted management-by-objectives systems, which required teachers to incorporate test results into their annual objectives. There was considerable variation among these districts regarding how sophisticated, comprehensive and tightly managed their regulation systems were.

Regulation systems can offer districts an opportunity to formalize the consultations we referred to earlier and to combine them with the judgmental innuendos of comparisons. One principal described his method for helping teachers establish their objectives as follows:

> I look at the scores to see where I should be concerned. I look by grade and by teacher to see which teachers have low scores. I have two conferences with each teacher, Fall and Spring, and in these conferences I tell the teachers each objective and we go over how they did and how it compares to the school average. [District 35]

This principal's description of how he works with his teachers in setting test objectives is remarkably similar to that of the principal quoted earlier who consulted with his teachers, but with one notable difference: the presence of formal objectives. And this difference seems to be very important. These district administrators had apparently successfully combined consultation and instigation, such that with very few exceptions teachers in regulating districts did take both the systems and their objectives seriously.

Administrators were aware that they had increased teacher anxiety. One district's evaluator told us they were "tightening up" their system, and said, "Now it may be scaring people. Now the goals are more sharply defined and we have changed our wording" [District 35]. Administrators often described their rationale for these systems with terms like "putting the heat on people" [District 35] The two districts with the strongest systems both carried the official line that the data would not be used to evaluate teachers, but unofficially, both administrators and teachers claimed the systems were moving in that direction. For instance, one administrator said, "Ultimately, we should be able to dismiss teachers and principals" [35]; and a teacher said, "A lot of people fear that it will be used to evaluate the teachers. We're not dumb, and we know it's headed in that direction" [19].

As administrators had hoped, teachers in these districts responded to these pressures by putting more emphasis on test scores. "I intend to teach the test," one teacher said. "I want the students to do well, and I will teach it to have my work look good. I'm not teaching anything that I disagree with, but the test does control my teaching" [19]. Another teacher described her method of transferring anxiety to students when she said, "The only way I've found to live with the test is to make it the most important thing in the kids' lives. I tell them over and over again that the test is the most important thing: 'Know that the test is all there is, and it will be on your report card.' I don't agree with this approach, but I'm a company person" [19]. As it turned out, teachers in this district often found they had to tie student grades to performance on the posttest in order to make sure students took the test seriously.

This attitude of conformance would no doubt be greeted happily by administrators, who for the most part hoped for, and assumed they were getting, this kind of response. But the desire to look good on test scores created a wide variety of responses other than those of conformance. These responses consisted of manipulating, in addition to instruction, test content, classroom composition, or test scores themselves.

Manipulation of test content, of course, can only occur if teachers participate in developing the tests. The results of their involvement were particularly apparent in district 19. There, for instance, the tests given to students in high school English courses are objective tests on grammar and sentence writing. They do not include literature, because teachers disagree on what literature should be included, and they do not include writing because teachers would be afraid to have an independent person score their own students' writing for them. In social studies, the high school test consists entirely of specific facts -- names and dates. The test includes no concepts, nor any analysis, because teachers didn't know how to write multiple-choice test items for more abstract content or more advanced skills, and could not agree on how to score essays. In both cases, the content tested has been limited to the most rudimentary forms of knowledge, not because these areas were judged to be more important, but because teachers could not develop test items of other areas. Even aside from debatable content areas, one teacher said, "We all end up writing items that can be answered by most of the students -- items that are not too difficult." Knowing the way in which test content was defined made compliance with the instructional goals implied by the test less attractive. The comments quoted earlier illustrating teachers' willingness to teach whatever is tested, despite their personal preferences, reflected the teacher response administrators had hoped to achieve. That is, the administrators expected teachers to be threatened by the systems and reluctant to comply, but they also expected that teachers would comply and that the test would drive the curriculum. But when teachers complained about tests driving the curriculum, they were not questioning the appropriateness, in principle at least, of having a uniform, explicitly stated curriculum. Rather, they were concerned about those things that were important and that should be learned, "but we can't test them, so they don't assume an essential role in the curriculum" [19].

In addition to examples in which teachers manipulated test content, we found teachers trying to manipulate the composition of their own classrooms. For instance, convinced that college-bound students generally did better than non-college-bound students, teachers fought over who got which students. And teachers in one secondary school fought over the prerequisites students needed for their classes. Since students did better in biology when they had already taken chemistry, biology teachers fought to get chemistry graduates in their classes. One teacher offered an explanation for these behaviors, saying: "No one makes allowances in the system for variations in student ability. They say that these variations will wash out, but anybody who has taught for very many years knows that there can be subtle differences in classes from year to year" [19].

The third way in which teachers tried to circumvent the system was by manipulating the test scores themselves. This was done in a number of ways. First, teachers could make a point of de-emphasizing the pretest to their students, encouraging them not to worry about it and not to take it too seriously, thus assuring themselves a lower starting point from which to show growth during the year [35]. Second, they can count their duller students as absent on posttest days, thus raising their posttest average and of course their gain for the year [35]. And, as one astute principal pointed out, "We can teach the upper [ability] kids well enough to bring up the low scores. At the administrative level, all they look at is averages" [35]. The most dramatic maneuver we found for manipulating posttest scores was a case in which some teachers broke into the principal's office and stole copies of their posttest from his safe [19].

These findings suggest that the extra emphasis given to tests under regulatory systems does indeed increase teachers' attention to test scores. But the concomitant anxiety, induced intentionally by administrators, leads to a variety of attempts to change not only instructional practices but also to change test content, class enrollment, and test scores independent of instruction. The administrators in these districts were aware of the anxiety they created, but, with the exception of the theft, were unaware of the many ways in which anxiety manifests itself. Rather than investigate the effects of the system, they tended instead to argue for its necessity. "No longer are teachers going to be left to practice in the absence of evaluation. People want to know how well teachers are doing, what the results are" [19].

EVALUATION

Incentives to pay attention to test scores move from regulation to evaluation when districts establish official policies to the effect that student test scores must be included in the formal professional evaluation of each teacher. This is the strongest method we encountered for stressing to teachers the importance of attending to test scores. Implicit in these strategies is the assumption that if teachers "used" test results, student performance on the test should increase. Hence, evidence of gains in test scores is even more important than merely evidence that the data were attended to in the teacher's planning. Now, not only must teachers use test information, they must use it effectively.

Two districts we visited included student test performance as part of the professional evaluation of teachers. Both of these districts were doing this for the first time the year we visited. One district also had a strong management-by-objectives system, and teachers had to set objectives that would be used in their own evaluation. The other district, rather than requiring verbally-stated objectives, required teachers to predict student performance on a posttest, and the prediction was later compared to actual student performance. In neither of these districts was the importance of these predicted scores or objectives in the overall evaluation of teachers made clear, either to administrators or to teachers.

Just as we found in districts using milder forms of inducements, there were teachers who accepted the fact of evaluation. One, for instance, said, "I didn't think it was unfair to judge us in part on the basis of test results. Some teachers don't mind it, but others complain and are worried about it. I have found myself teaching to the test -- you know we are told we are going to be tested on it, so it is only natural that we prepare kids for the test. I feel that as a teacher, you are supposed to do a job, and they are only telling you what to do" [7].

In both districts, it was the principal who would eventually evaluate teachers. Teachers either had to submit their objectives to their principal for approval, or had to obtain their principal's approval for their predicted scores. One principal told us how he appealed to his teachers' experience with him so they would trust his judgment.

> The folks in this building have been able to size me up. Initially the new state regulations on observation made them kind of nervous. They seemed concerned about what would happen to the relationship between me and them, what would our interactions be like. They were concerned about what the relationship would be like with the superintendent, but I reminded them that I wasn't going to be any different. I suggested that they think over their careers in the schools and see that things have been pretty stable. I said the person in the room will be me and the kind of things I do in evaluation will be the same as the kinds of things I've said before. [District 7]

A high school teacher in the same district reflected on the importance of his relationship with his principal: "I'm not sure how someone else would use the information they have on us for evaluation. I mean, like the test scores. I have a decent supervisor now, but what if I get a new one? I don't know who the next supervisor will be, and I don't know how he would use the stuff" [7].

Despite the strength of some of these personal relationships, administrators and teachers in both districts were aware that a great deal of pressure was now being placed on teachers. Yet, despite the pressure, we did not find the behavioral manifestations of anxiety that we found in districts with regulatory control. Instead, teachers apparently resigned themselves to the evaluations, and responded by turning apathetic or cynical, and many discussed the possibility of leaving the profession:

> I don't know if [teacher morale] goes into the negative numbers or if it ends at zero. There's lots of moaning. People know here it's either sink or swim. Sixty to seventy percent of people in this school want to leave. But there's just no place to go. There are no teaching jobs. [27]

> The big question facing me is, do I really want to stay in the profession. There are a lot of people I know who are really good teachers who are not coming back. One thing I've liked as a teacher here was that the stress factor wasn't so great, but now, if they are going to increase the stress without increasing the pay, I may decide to go back into business. [7]

> It causes cheating and competition. But I think we'll see more testing because of accountability and back to basics. More and more and more testing. [27]

And despite the strength of some personal relationships between principals and teachers, cynicism often was not squelched by principals, but rather spread to them. One principal in the district where test score predictions were required, for instance, put his tongue in his cheek when he said, "Everybody did make a prediction, ... but I didn't realize how little we expected kids to know" [7].

These attitudes are probably due in large part to the sense of helplessness at being able to control the outcomes of their own evaluations. A teacher who had to predict scores, for instance, said:

> It's a joke. How can we really anticipate how these new, strange kids will perform? This is a Catch-22. If you predict too low they have you: they say you don't expect much. If you predict too high, then you're a loser. On top of this, some teachers teach to the test. What kind of instruction is this? [7]

In the other district, one teacher showed us her objectives, saying "These aren't my goals. I just copied the sample, like everyone else. . . . I put these on paper because I had to" [27].

Evaluation as a strategy for increasing teachers' attention to test performance is the ultimate form of stress that we observed. Though we found some oblique references to attempts to undermine the system, they weren't as frequent or as serious as those we found under regulatory systems. Instead, teachers seemed to have apathetically resigned themselves to their fate, and contemplated leaving more than subversive attempts to survive.

CONCLUSION

Our data are not so complete that we can speak to the relative popularity of any of these administrative methods, nor can we indicate how likely any of these various kinds of teacher responses are. Our purpose here was merely to indicate the range of responses that can and do occur. Many of the responses we observed were counterproductive, and the data suggest that the frequency and diversity of counterproductive responses increases as stress increases. But to leave our conclusion at that would be overly simple, for it begs the question of how many teachers respond counterproductively, as opposed to productively, to any given emphasis. The correlation between stress on tests and stress in teachers is not perfect: some teachers may experience considerable anxiety even with relatively mild inducements; others may weather even the strongest pressures easily. The problem of evaluating the overall effects -- and effectiveness -- of administrative strategies such as those described here is a big one. The strategies we have described here are not easily isolated from other functions of the school district. They are part of the very fabric of its operations. They are not the result of particular decisions; they are the decision-making strategies themselves.

The evidence presented here suggests that the assumptions underlying these systems may not be correct. The first set of assumptions led to the belief that these systems would create incentives for rational instructional decision making. This would occur in three ways: requirements to use the test data would automatically lead to rational uses of the data, specification of the goals of education would make teachers more goal oriented, and the formalization of individual responsibilities would motivate teachers to improve their progress toward those goals. The four types of systems reviewed here differ in the extent to which they meet these assumptions, but taken together they suggest that enforced reliance on tests does not necessarily motivate greater rationality in instructional decision making, but instead motivates staff to increase the test scores themselves, regardless of whether student knowledge is changed in the process. Teachers are motivated to manipulate all the variables that influence test scores -- instruction, class composition, test content, and so on. The second set of assumptions led to the belief that teachers were capable of using test data to make rational instructional decisions. Teachers were assumed to be capable of this because they would know how to interpret test score patterns and identify the knowledge their students were lacking, and because they would be able, given that knowledge, to modify their instruction to provide students with that knowledge. The fact that teachers chose to increase test scores by means of non-instructional manipulation indicates that these assumptions may not be correct, for if teachers could readily interpret the evidence and modify their instruction accordingly, they would have less need to manipulate other variables that influence test scores. At bottom, these systems err in confusing the tools of rationality with rationality itself. Rather than tests becoming servants to rational decision makers, the decision makers became servants to the test.

THE ROLE OF THE IN-HOUSE EVALUATOR

Evaluation is an inherently contradictory activity. Evaluators are expected to facilitate change, yet clients resist change. Evaluators are expected to help organizations achieve their goals, yet organizations may consist of parts whose goals are incompatible, so that helping one group entails hindering another. Evaluators are expected to produce decision-oriented information, yet clients can rarely identify decision options far enough in advance that they can be studied. Evaluators are often expected to observe organizational activities from an objective position, yet their credibility may depend on being perceived as sympathetic friends. Most of these tensions are inherent in the task of evaluation.

Contemporary literature on the role of evaluation approaches these problems from two perspectives. On one side are articles pointing out the difficulties that may face evaluators. For instance, Weiss (1973) points to such organizational constraints as conflicting perceptions of the purposes of evaluation and high staff turnover; Cohen (1970) points to the multiple motives that promote programmatic decisions, and Lindblom (1965) illustrates the inherently political nature of decision making processes. On the other side are articles suggesting methods evaluators can use to accommodate problems such as these. Wise (1980) proposes that evaluators adopt the role of teachers, Zeigenfuss and Lasky (1980) propose a management-consulting role, Krathwohl (1980) suggests the need for a negotiation facilitator - fact finder role, and Barkdoll (1980) distinguishes three evaluator roles: investigators reporting, highly technical analyses, and consultant-consensus building strategies. Several authors have extolled the virtues of social experimentation as the best method for discovering the real effectiveness of social programs (Bennet and Lumsdaine, 1975; Boruch and Reicken, 1975). These two bodies of literature play in tandem. One group points out problems and the other offers possible solutions to these problems. Together, they have not resolved the contradictions inherent in the role of evaluation, but they have made considerable progress in clarifying these dilemmas.

But much of the literature on evaluation and its function is based on assumptions that have not been explicitely reviewed. For instance, the literature on evaluation often assumes that evaluation is carried out in relationship to a discrete conceptual entity -- a demonstration program, or a particular policy. Yet many clients may desire evaluators to inform them of their general state of affairs. They may ask for statistical indicators of their organization's overall activities or of its environment, and expect evaluators to help them interpret these indicators, and they may ask for these services without defining any particular purpose for the investigation. No conceptual entity is under particular scrutiny, no particular problem is awaiting a solution. Second, the literature often assumes that the evaluator is an independent consultant, a freelancer called in on temporary assignment to evaluate this conceptual entity. Yet many evaluators are in-house evaluators. They are not on temporary assignment, and if they are to remain members of the organization, they must find ways of responding to unfocused inquiries as well as to focused inquiries. Finally, most of the evaluation literature assumes that one of the evaluator's first responsibilities is to avoid being compromised by the organizational and political conditions that appear to be inherent in decisions and in decision making processes. Whether evaluators can in fact be completely objective is a subject of debate (Cooley, 1980;

Kean, 1980), but the general preference for neutrality is evident in nearly all proposed evaluation roles. The most important part of this assumption is not the assumption that neutrality is important, but the assumption that the evaluator can control his or her own neutrality. It is possible that the decision-making processes and the organizational dynamics within the client organization are so powerful that the evaluator cannot operate independent of them.

This paper is about in-house evaluators. In-house evaluators are permanent members of their organizations, and their job is to observe and assess the activities of the organizations to which they belong. The three assumptions mentioned above are particularly problematic for in-house evaluators because the permanence of their positions may depend on factors other than satisfactory performance of the role assumed in the literature. For instance, they must learn to be helpful in interpreting indicators when the client's inquiry lacks a clear focus. Furthermore, even when the client wants an evaluation of a specific conceptual entity, such as a program or a policy, both the in-house evaluator and the members of the organization as a whole know that that entity is completely confounded with a performance entity -- the people or administrative divisions who operate the program in question or enforce the policy in question. The confounding of conceptual and performance entities automatically places the evaluator in an adversarial position relative to the program being evaluated. If the in-house evaluator plans to remain in the organization, he or she must develop an organizational role that not only accommodates those responsibilities that the evaluator feels are professionally important, but that can also accommodate those organizational responsibilities that other members of the organization feel are important. And responsibilities may be such that they do more than merely compromise the evaluator's role -- they may in fact define it for him.

In this paper, we describe the activities of in-house evaluators in 16 school districts. Each district has a reputation as a place in which evaluation and test data are used well. From a pool of such noted districts, these were chosen to form a sample that would be diverse in geographic location, wealth, size and ethnic balance of the student body, and in the apparent activities of the evaluation office. The data gathered from these districts were entirely qualitative, coming from interviews and observations. In all districts, the focus of data collection was on the use of information more than on the role of the evaluator per se, but evidence regarding the evaluators' roles was a natural byproduct of this line of inquiry.

Our intent is to shed light on the relationship between in-house evaluators and their organizations. The available literature has tended to de-emphasize the role the organization has in determining the role of the evaluator, but this paper focuses on that influence. The paper has three parts. The first describes several ways in which organizations define their evaluators' activities. The second describes three examples of evaluation units which had not been able to adapt to the organizational roles that were expected of them. The third describes the roles that the evaluators or evaluation units in the remaining districts had adopted, and shows how these roles fit into their organizations.

ORGANIZATIONAL INFLUENCES ON THE EVALUATOR'S ROLE

Much of the literature on what evaluators ought to do appears to have been written on the assumption that the evaluator's role is something he or she can choose. Yet the organization exerts a great deal of influence over the role of the in-house evaluator, for evaluators are hired to fulfill specific organizational needs and their budgets and sizes are determined by the school board's estimate of what is required to meet those needs. In-house evaluators compete for funds with the very programs they evaluate. Under these circumstances, the organization necessarily influences at least some aspects of the evaluator's activities.

The most obvious influence the organization has on the evaluator is in determining whom the evaluator will serve and what services will be provided. Though evaluators have some discretion in these matters, they are usually hired to fulfill pre-defined needs. In all 16 of these school districts, the two activities that consumed most of the evaluators' budgets -- the testing programs and the evaluations that were mandated by state or federal funding agencies -- were prescribed.

All school districts had at least one achievement testing program. Some had more than one, combining, for instance, a nationally standardized norm-referenced test with a locally developed curriculum- or criterion-referenced test, a state assessment, or a state-mandated competency test. Generally speaking, districts engaged in multiple testing programs, with the evaluation department administering the norm-referenced test and the curriculum department administering the curriculum- or criterion-referenced test.[1] State mandated tests followed no clear pattern. Rarely did in-house evaluators have much influence in determining the content of these various tests. In these 16 districts, only three in-house evaluators -- one who was labeled an evaluator and two who were titled curriculum directors -- determined test content. The content of state assessments and of state-mandated competency tests were usually dictated by the state, and the remaining tests were either chosen or developed by committees of teachers or occasionally by the superintendent, though in-house evaluators often sat on test selection committees.

In-house evaluators also had little discretion over the timing of the test administration or the format in which the data were released to various audiences. Locally-developed tests were usually part of management systems, so their administration and the dissemination of their results were dictated by the needs of the management system. Purchased tests were generally administered according to a district policy stating the grade levels to be tested and the time of year testing would occur. The computation of test scores was either done by the publisher or it was done in-house using the publisher's procedures. The results of purchased tests were distributed in one or more of the following formats:

- Gummed labels for each child, stating his or her score on each subtest. These were added to other gummed labels in each child's cummulative file.
- Slips of paper summarizing each child's scores on each subtest. These were sent home to parents.
- Printouts of each class's performance. These were given to each teacher. Their content varied considerable, and could include lists of outcomes by child, by subtest or by item as well as various kinds of patterns of outcomes within the class.
- Printouts of the performance of each school building, providing breakdowns by classroom, grade level, subtest, and so on. These were given to school building principals. In some districts, principals also received a copy of each teacher's printout.

- A district-wide summary, usually in book form, presenting district-wide averages and one- or two-page summaries of each building in the district. These were usually printed for general distribution.

Though the in-house evaluator could determine the format of the last item listed above, the formats for the first four were generally determined by the combined forces of the district budget and the publisher's options. Publishers offer a variety of analytic options and districts purchase those they can afford. Not surprisingly, in-house evaluators spend more time thinking about ways to improve the fifth item than they do ways of improving the first four.[2]

Mandated evaluations are not as overtly prescribed as are tests, but they are nevertheless prescribed, usually by a combination of the following: the program regulations; the design of the district's own testing program; local traditions; and the sheer volume of mandated evaluations that are produced. Program regulations define the kind of information that must be provided and often prescribe the format for providing that information. For instance, one program may require information about student achievement relative to a norm or a comparison group, while another requires it relative to program goals. The district's own testing program constrains the evaluator's options even further. Most districts cannot afford, and do not want to burden their students with, additional testing done solely for the purposes of mandated evaluations; consequently programs are evaluated with designs that conform to the extant district testing program. Once an evaluation design has been developed to accommodate both program regulations and the district's testing program, it may simply be repeated year after year, until tradition dictates that it not be changed. Finally, in some districts, the volume of mandated evaluations that must be completed each year forces the in-house evaluator to routinize his or her evaluation procedures. The combination of program requirements, local testing routines, tradition and volume of mandated evaluations contributed in many districts to the mass-production of mandated evaluations. They were not projects that were undertaken individually, but were rather things that were administered, in much the same way that testing programs were administered.

In addition to prescribing the procedures by which these mandated evaluation activities would be carried out, school districts generally determined who evaluation units would serve. Decision-making in school districts can occur at several levels of the organization: in the classroom, in the school building, in centralized offices that operate programs or develop curricula, by the superintendent or by the school board. Which audience received the benefits of the in-house evaluator's services depended in large part on the title and organizational location of the evaluation unit. In these 16 districts, eight in-house evaluators reported directly to the superintendent of the district, three to an assistant superintendent for instruction, two to assistant superintendents for planning and budgeting, two to assistant superintendents for federal programs and one to an assistant superintendent for administration and personnel. And the titles given to the evaluation units further indicated the unit's mission. In addition to such terms as "research", "evaluation", or "testing", in-house evaluation units had these terms in their titles: curriculum (2), instruction (1), policy (1), planning (2), and accountability (1). Those units whose titles or reporting chains had to do with curriculum and instruction were more likely to serve school buildings and to spend their time interpreting patterns of test scores. They tended to engage only in summative evaluations that were mandated. Those whose titles or reporting chains had to do with planning and accountability were more likely to serve senior administrators or board members, to conduct summative evaluations, and to engage in ad-hoc policy analyses.

Finally, in addition to these organizational influences on the role in-house evaluators could play, the evaluation role was defined by the character of the organization itself -- its standard operating procedures and its management and decision-making practices. Each school district had its own style of operation and its own pattern of relationships among its members. These patterns and procedures constituted the organization's problem-solving style, and they determined who would need data as well as what kind of information they would need and when they would need it. Some districts emphasized their conceptual entities -- bilingual education, special education, vocational education, and so on -- while others emphasized performance units -- classrooms and school buildings. In the first kind of organization, principals and teachers who served diverse populations of children could report to two or three program directors, whereas in the second teachers were responsible solely to their principals and principals solely to the assistant superintendents directly above them. In the first kind, policy makers were concerned about the coordination among programs and about the effectiveness of programs, whereas in the second they were concerned about the quality of instructional oversight procedures and with performance of individual teachers and principals. In the first, evaluators were often expected to provide summative evaluations of programs and to document the overlap and coordination among them, and in the second they were expected to provide accountability data to each supervisor-subordinate dyad in the instructional oversight system. Yet a third kind of organization tended to emphasize neither conceptual units nor performance of individuals, but instead devoted its attention to monitoring indicators of district-wide performance. When problemmatic data appeared, solutions could be sought through either conceptual or performance units.

These, then, are the organization's influences on the evaluator's role. First, the organization defines the size of the evaluation office and determines its budget. Second, it defines the largest part of the work -- the testing program and the mandated evaluations. Third, it defines the subject matter with which evaluation will deal and the audiences for any work done over and above the two primary activities of testing and mandated evaluations. And finally, it defines the parameters of the work by emphasizing a particular problem solving style. These four organizational determinants of the evaluator's role are not independent. Problem solving styles entail assumptions about causal relationships, about how various organizational efforts eventually influence educational outcomes, about which things are likely to be responsible for problems, about where to look for solutions to problems and about who should be responsible for correcting problems. The organization's problem-solving style includes a rationale for why evaluation and test data are needed and assumptions about how these data should contribute to problem solving. That rationale in turn dictates the title, organizational location, size, budget, primary activities and primary audiences for evaluation units -- the kind of information it will produce and the manner in which it will present that information to the organization. It is this complex web of interconnected behaviors and assumptions, then, that constitute the organizational context within which each in-house evaluator must work. Though evaluators may be able to adapt relatively easily to the substantive needs of whatever clients are assigned to them, and may be able to adapt to whatever resources they are given, their most significant adaptations will be to the organization's problem-solving style, for that adaptation will determine how well they will be able to serve the organization.

PROBLEMS OF ADAPTATION

The importance of adapting to the organization's problem-solving style is particularly apparent when observing school districts which have recently experience radical changes in their problem-solving strategies. Such changes can occur when, for instance, changes occur in the budget or in the balance of power, for these can create tensions which in turn lead to less uniform or less predictable behavior among the members of the district. When the conventional strategies are abandoned, the role of the evaluator is no longer clear, the evaluator becomes frustrated, and other members of the district are confused over what the evaluation unit should be doing and disappointed with what it has been doing. Such changes occurred in three of the districts participating in this study. Districts 83 and 72[3] had had rapid changes in superintendents, and both had undergone one school year during which three different superintendents held office. District 18 had also had turnovers in leadership, though not as many, and it had experienced other kinds of setbacks as well. It served a single-industry community and that industry had lost a great deal of business. The community was pushed into a severe economic depression, and the budgets of the city and the school district were affected.

District 83 had been decentralized for several years, and its leadership turnovers were in part the result of political difficulties inherent in trying to re-centralize decision-making and curriculum control. The evaluation function was one of several that became centralized during this tumultuous period. Under decentralization, the evaluators had been assigned to clusters of school buildings, and were housed outside the central office, near their clients. Their role had been that of consultant. They helped their clients find patterns in test scores and other data, and helped them interpret those patterns. Each evaluator was closely identified with her or his clients. Under centralization, these evaluators were brought together and housed in the central administration offices. One of them was chosen to direct the new unit. An assistant superintendent explained the changes like this:

> The old view of evaluation was that it should work directly with teachers and parents, helping them in the process of understanding their children's test scores. However, the current administration has no belief in process, only in outcomes, and therefore doesn't value the close working relationships that the evaluators had worked out with schools and teachers and doesn't believe they should be spending as much time as they were actually going out into the schools. The new superintendent wants answers to questions like, "Is such-and-such working?"

This new evaluation purpose suggested a new organizational problem-solving style, and it was not completely understood or accepted by all administrators or board members. Several complained about the re-organization, claiming evaluators could no longer serve schools as well as they had in the past. Others complained that the unit had been and continued to be ineffective in stimulating instructional change. One board member espoused multiple views on what evaluation should do within a single interview, unaware that she was contradicting herself. Her remarks include these:

- Evaluators should be advisors to instruction
- Evaluation information should be used by practitioners
- Data such as test scores should be used for making curriculum decisions.

and these:

- The evaluator's prime function should be to supply information to the central administration. They need data.
- Evaluation should be separate and on-going, like in businesses
- We need on-going evaluations of programs; otherwise you don't know when to put them out of business
- The current unit spends a lot of time interpreting test scores. That's necessary but its not sufficient.

The new evaluation director had not been a supervisor before and many of his staff resented his position, feeling they were his equals rather than his subordinates and that their allegiances belonged to their former clients. He described two problems he had to solve. One was to develop a method of interacting with and reporting to the superintendent, and the other was to find a way to "abolish the old loyalties and establish new ones." As far as the work of the unit, the director saw no change in its purpose. When asked who his primary clients were, he said, "teachers." He perceived centralization as stemming from a desire to unify the curriculum by unifying the kinds of analyses and consultation that were offered to school buildings. He did not perceive a change in organizational problem-solving style, nor did it occur to him that the unit might have been created to serve senior administrators rather than school building staff. In fact, he resisted the requests he had received from these potential new clients, sensing that the data they asked for might be detrimental to former clients. He told us of pressures to report test scores broken down by school building and by ethnicity of students within school buildings. He said he had tried to present the results across schools in a way that made "direct comparisions between schools as difficult as possible," and that he had refused to break school building scores down by student ethnicity on the grounds that such data could be interpreted "either as a failure to meet minority needs or as indicative of minority inferiority or both." One associate superintendent, after discussing the problems of the newly centralized evaluation unit, then moved on to discuss another administrator who was not adapting to changes in the district. With regard to the other administrator, he said, "I'm worried about her survival too" [emphasis added], thus indicating a suspicion that the evaluation unit, or its new director, may not survive this transition.

The fate of the evaluation unit in district 72 has already been sealed. This unit also began by bringing together evaluators who had formerly been dispersed throughout the system. The new unit got off to a better start than did 83's new unit, in part because the superintendent who established it hired a new evaluator to operate it, someone who knew and sympathesized with the superintendent's views of evaluation. Both he and the superintendent had a clear vision of the organizational problem-solving style the district should have and about the role that the evaluation unit should play. That superintendent was fired, and the evaluator survived two more superintendents before his entire unit was abolished. The director of the unit was not too sorry to leave, and said he had not accomplished what he had hoped to do. "There's still no locus or system for information or for its use. Information is both politically and technically dispersed." Since this was this evaluator's only experience in a school district, he assumed the problems he encountered were endemic to all school districts. He said he thought the demise of his evaluation unit was:

> Partly due to the economics and politics of education and the lack of vision that accompanies them. No one thinks in terms of long-range planning -- they only think of piecing things together for the moment. But that's also because education -- school systems -- are too unstable for planning. The contingencies make it difficult to have a coherent plan for anything. . . . There are just too many uncontrollables. Something could fall off the table at any moment.

This evaluator's disillusionment was only partly due to the traumas his district was undergoing during his tenure there. It was also partly due to his own idealistic notions about what his role should be. Rather than adapting his services to the organization, he expected the organization to adapt to the services he wanted to provide.

The evaluation unit is district 18 was experiencing a different kind of problem. The district had had three superintendents in the past five years and it suffered from severe budgetary problems. Thoughout all of these organizational traumas, the evaluation unit had not been reorganized. But the district's problem-solving strategy had changed drastically. The current superintendent did not rely on a cabinet, nor did he request advice from other staff, from committees, or from the evaluator. Nearly all organizational problems were solved by the superintendent with the help of one confidant. Changes were announced to the rest of the organization without discussion or rationale. This secretive behavior lead other members of the district to feel there was no rationale for decisions, no problem-solving style that they could comprehend. They suspected every action of being motivated by a hidden agenda and suspected their colleagues of secretly influencing decisions. The director of evaluation was doubly injured by this problem-solving style. He was as ignorant as anyone else about how and why decisions were made, and consequently he had no idea how he could have contributed to them. In addition, he was new to his position, and unlike his predecessor, who had become very attuned to the political climate and strategy, this evaluator had been sheltered and now had no idea how to enter the inner circle. Like others in the district, he complained bitterly about being left out, and like others he blamed his difficulties on others. In his case, he felt his supervisor had not worked hard enough generating support and enthusiasm for his unit.

These three evaluators were the victims of changing circumstances. All three had theories of how evaluation should contribute to problem solving, and two of them had had experiences in which their visions had been realized. But all of them were, at the time this study was being conducted, in situations that did not mesh with the roles they wanted to adopt, and the strain that resulted had led to antagonisms, disappointments, and frustrations, not only for the evaluators but for their clients as well.

EFFECTS OF SUCCESSFUL ADAPTATION

The survival of the in-house evaluator depends heavily on his or her ability to develop an evaluation role that blends with the organization's problem-solving conventions. Since school districts vary in their organizational strategies, the role their evaluators adopt must also vary. Reports indicating each classroom's performance, for instance, will not be useful to the school district which is organized around conceptual entities and which makes its major decisions with regard to those entities, and experimental comparisons of conceptual entities will be of little value to a district which, though containing such entities for funding purposes, has in fact decentralized its decision making responsibilities to the individual classrooms. The three evaluators described in the preceding section had adapted to one organizational style and were unable to adapt to another. The evaluators in the remaining 13 school districts had developed roles that successfully blended into their districts and their roles can be grouped into four broad categories.

- The Technician Three evaluators did little more than administer the local testing program and conduct mandated evaluations. If other activities were undertaken, they were only done when specifically requested. The evaluators made no attempt to assist their clients in any ways other than by producing the data.
- The Participant Four evaluators (in districts 4[4], 25, 50 and 220) took a genuine interest in the issues that faced their clients and worked closely with them in attempts to understand their problems and to help solve them.
- The Management Facilitator Evaluators in four districts (districts 17, 19, 27 and 35) viewed themselves as part of the management team and as having responsibilities primarily related to making building-level staff more accountable.
- The Independent Observer Two evaluators (in districts 57 and 115) adopted the role of neutral evaluators. Though they tended to serve some audiences more than others, they did not sympathize with their clients, nor did they participate in problem solving beyond the provision of information.

The role of technician is sufficiently limited as to be of little use in furthering understanding of the working relationship between in-house evaluators and their organizations. Consequently the three examples of this role will not be discussed in this paper. The other three roles, however, each present particular kinds of problems which we might profit from seeing. The role of participant and of neutral observer, for instance, reflect two very different, and in many ways opposite, ways of handling the evaluator's competing obligations to provide assistance and to be objective. Whereas the participant may seek handy rules of thumb, the observer is more interested in ascertaining absolute and irrefutable truths. Whereas the participant's role is more likely to be one of helping others understand their situation in general, the observer's role is more likely to be one of passing judgment on the effectiveness of different organizational activities. Whereas the participant provides information for particular _people_, the observer provides information for particular _issues_. The second role, that of management facilitator, falls somewhere in between these two extremes. The management facilitator is a participant in the sense that he or she helps managers in their work, but is an observer in the sense that he or she measures the effectiveness of the performance of the manager's subordinates. These three roles are not distinct in the sense that an evaluator who adopts one role never adopts any other roles; in fact, nearly all of the evaluators participating in this study modified their roles under certain circumstances. But the roles defined here do describe the predominant tendencies of these evaluators, and in so doing, they highlight the problems as well as the advantages of each kind of role.

The Participants

Participants can be characterized by two important features. First, they tend to serve people rather than issues, and second they tend to take their clients' interests as their own, and to have a relatively complete understanding not only of the substantive issues facing their clients but also of how their clients perceive and think about those issues. Two evaluators fell into this category because their primary responsibilities were not, in fact, in evaluation. They were educators first, evaluators second. The evaluator in district 4 was a curriculum director who also administered the district's testing programs and consulted with building staff on the meaning of test scores. The evaluator in district 25 had been temporarily assigned to be the Title I evaluator, but was reassigned to another program position two years later. These two evaluators were consumers, as well as producers, of evaluation data, and their conversations with other members of the district were not interactions between evaluators and their clients but interactions among educators.

When asked about his role as evaluator, the curriculum director in district 4 claimed to use test data as a conversation starter. He described his conversations with school principals as follows.

> I try to get them to generate usable questions. I'll ask them, "Why do some kids in the fourth grade in one school have a reading mean of 58.6 . . . while they have a math mean of 87.6? Why is there such a discrepancy between reading and math scores?" And then I'll have them look at the distribution of teaching time in the different subjects and see if they're giving students more time for math than they are for English.

Once he got a conversation going, however, it extended far beyond the test data that initiated it. One principal described his interactions with the curriculum director like this:

> I ask him for information all the time, and vice versa, especially to bounce insights off him, and especially about writing. I'm used to operating on feel; he has not only that but can pull up the structure to back it up. When it comes to writing, he and I have different assumptions about how to teach it. My basic premise is that kids will learn to write better and more quickly if they do their own editing. His premise is that it doesn't matter who manages the editing as long as the kids do some revisions.

The woman who was temporarily assigned to the position of Title I director had a similar working relationship with the Title I program director. She was not his advisor but his partner. They made decisions together and she had as much interest in the program as the director did. When she described one recent change they had made in one of their Title I programs, she said,

> I'm keeping my fingers crossed that the new program will show better results. The supervisors say it is going well and that the kids are on schedule, so I'm hoping the results will show up on test scores.

On one occasion the evaluator joined the director and some supervisors in visiting schools to see why a program wasn't doing as well on the evaluation as they thought it should. Among other things, they discovered that the aides in the program were serving up to 60 children apiece. After some discussion they decided to concentrate aides so that each would work with only 10 children per year. When the evaluator wrote her end-of-year report, she included this finding regarding aides, and included a recommendation that aides be concentrated on 10 children apiece. The report did not convey new information to the director, nor did it present a recommendation from an evaluator to a client. Rather, it was a written record of the problem-solving interaction that had occurred between these two concerned individuals.

The evaluators in the other two districts, though they behaved as participants, had permanent positions as evaluators and were located in units organizationally separate from their clients. Yet, although their titles and locations suggested they had a distinct purpose, they strongly identified with their clients.

In district 220, this identification occurred despite considerable rhetoric to the effect that evaluators should be independent and objective. Program directors claimed the virtues of independent evaluators with comments such as this:

> They are in evaluation. We decided early on that they should be independent of me. But they are a service to me, and I provide their salaries.

And evaluators in district 220 claimed to value their independence when they made comments like this:

> When you have an evaluation person attached to program staff, they don't have enough clout to stand up to the program director, and they often become administrative assistants. You lose the ability to stand up and give your findings.

Yet despite their apparent belief in independence, these evaluators were quite attached to their clients. The evaluator quoted above, for instance, spent the first 40 minutes of his first interview with us providing programmatic background, a behavior we frequently encountered with program directors, but only very rarely with evaluators. His interest in "his" programs was also apparent in the tenor of his remarks -- comments like, "These are mind-boggling problems and the [teacher's] union doesn't think about them," and prideful comments about the quality of the programs. His services to one of these programs were particularly appreciated by its director, who equated the benefits of evaluation with the growth of her program:

> I have the highest regard for evaluation. We've been able to use all the evaluations they've provided -- the program has grown from 16 schools to 77 . . .

This director also showed us an example of one of her evaluator's products. It was a one-page summary of the program's effectiveness, a bar chart comparing her program with a competing program, and her program's bars were longer. The sheet did not mention that the comparison program served only children who had low test scores, whereas this director's program had no such admission requirement. The persuasive bar chart was apparently produced by an enthusiastic evaluator who forgot about these important pre-program differences among children

Another evaluator in this same district served the superintendent, a man particularly interested in using test data to assess the callibre of school building personnel. The superintendent described his interaction with the evaluator like this:

> We look at scores at the district level. Then we look for trends in the district and we look for problem areas. Once we find the problem area, we identify the schools that have the most problems, and right now we are looking at particular teachers.

The evaluator had a great deal of respect for the superintendent, both as a manager and as a person who understood and used data. She had adopted her superintendent's outlook, and described it this way:

> We can see differences from school to school, and it is so obvious that there is an administrator effect that makes a difference -- by administration, I mean it could be a principal or a reading coordinator, but their effect shows up. I've seen enough buildings to tell. [The superintendent] works hard to weed out the dogs.

Although the relationship between these two was ostensibly one of advisor and client, it is difficult to say who was influencing whom when it came to interpreting test scores. For instance, although the superintendent had said the evaluator was helping him use test scores to look for problems and to identify weak staff, the evaluator often interpreted the test data using the superintendent's point of view, rather than offering the superintendent an alternative point of view. In reference to an elementary school principal who was favored by the superintendent, for instance, she said,

> [He] is an example where he is a good principal but the test scores don't show it. He has lost so many good teachers, and there are just so many variables that affect test scores.

District 220's evaluators, then, although located apart from their clients in order to maintain neutrality, have in fact adopted their clients interests as their own. Not only do they provide information that fits their clients' information needs, but they interpret the evidence in ways that reflect their clients' points of view.

The fourth evaluation unit that participated in problem solving was in district 50. District 50 also had an organizationally separate evaluation office, but the unit did not divide its staff among several clients as district 220's unit did. Instead, the problems facing the senior administrators -- declining enrollments, rapid influx of non-English speaking students, and threats of desegregation law suits -- were so urgent that they enveloped all evaluators as well as all administrators. The evaluators played a particularly large role in planning for desegregation, and their contributions were well received. When they were asked to account for their success, they attributed it to the fact that they produced working drafts, rather than finished reports, so that planners had information when they needed it most. However, the assistant superintendent who was officially in charge of the planning team referred to the evaluators' contributions by saying:

> They were the people with the professional planning background so we drew heavily on their background, especially in terms of procedures and technical skills. They provided the leadership.

On the surface, the evaluators in these four districts appear to be quite different. One was a director of curriculum, another a Title I evaluator, another a large evaluation office serving several different groups, and the fourth an office serving primarily one group. Yet in practice they were very much the same. They took their clients' problems as their own, and became genuine partners in planning. And in each case, their role fit well into the organization's problem-solving style, for their clients tended to engage more in incremental, trial-and-error adjustments to practice than in major decisions about major service components. The advantage of the participant role is that, since clients often do not approach evaluation data with particular questions in mind, the evaluator can assist analysis and interpretation better by becoming almost an alter-ego for their client. But the disadvantage is that, in so doing, evaluators automatically abandon neutrality and adopt their clients' interests as their own. They cannot anticipate their clients' information needs without becoming alter-egos, and they cannot become alter-egos without taking the same interest in their clients' responsibilities as the clients themselves take.

The Management Facilitators

The role of management facilitator offers a unique blend of participant and observer. The four evaluators who adopted this role were like participants in that they had embraced a particular client's point of view, but they differed from participants in two important ways. First, they perceived their data as authoritative in the same way that observers tend to do, and consequently they were less likely than their true participant colleagues to rely on non-evidential considerations in their assessment of current affairs; and second, they did not tend to assist clients by helping them sort out evidence and interpret its patterns. Rather they provided evidence whose interpretation was already assumed. The management point of view was that subordinates -- school building principals and teachers -- were responsible for the outcomes measured, and if the outcomes were not satisfactory, those were the individuals responsible for changing them.

Two evaluators fell into this category because they were in fact managers. The data they gathered and analyzed were for their own supervisory purposes, though in both cases they also shared the data with school principals and teachers so that these subordinates could use the data to develop objectives or improvement plans. The supervisor in district 19 described his system this way:

> The system provides a way to control what the teacher teaches. If they don't teach the objectives, it will become abundantly clear. Teachers know what will be tested in every CRT module. They don't get the post-test itself, but they get one that is comparable, having been constructed from items that were in the item bank.

And in district 27, the assistant superintendent described the effect of her new data-based management system on principals like this:

> They were overwhelmed. Everyone was very comfortable here before. . . . We uprooted the whole system, and any time there's change there's stress. We had a lot of change and a lot of stress. But we gave them assistance too. Eventually the principals saw they were producing more and they were pleased with it.

Both of these supervisors were more interested in supervision than in evaluation, and their supervisory interests influenced all their data collection and reporting decisions. For instance, the supervisor in district 27 described an instrument she had developed for principals to complete. The last page of the instrument was what she called a "B.S. page." This page had questions about the value of the instrument itself and about whether it was well constructed. But she did not attend to anything principals wrote on this page. She included the page because it gave principals a sense that "their individual beefs would be heard." In fact, she believed that a "good principal" would not fill out the B.S. page because "they wouldn't need the extra pats on the back or the extra opportunities to express grievances." The format she used to disseminate data was also based on only her own point of view. One principal revised all the printouts he received before sharing the data with teachers. His reason:

> I don't want to use it this way, for fear of intimidating teachers. This [printout] looks like a way of tracking teachers, not a way of understanding the needs of kids.

And it is.

The remaining two districts had organizationally separate evaluation offices, but the work done by these offices was designed primarily to facilitate similar management practices.

District 17 was just beginning to strengthen its central management system at the time of our visits. The district was a consolidation of several smaller districts, and for years the central management had permitted regional autonomy as politically necessary to the survival of the consolidated district. At the time of our visits, senior administrators were increasing their authority over schools by relying on a new state mandate which required districts to inform their communities of their objectives and to inform parents of how well students were doing on those objectives. Both district 17's administrators and its evaluators interpreted the mandate as requiring a uniform curriculum and a system for tracking progress of each student and each classroom through the curriculum. Administrators interpreted the law in this way because they had wanted to centralize decision-making anyway, and to make schools less autonomous. The evaluators interpreted the mandate this way because they had already developed a large bank of test items catalogued according to grade level and objective, but had not been able to convince teachers to use these items. They wanted their system to have more influence on instruction than it currently had, and they complained about teachers in their district who saw no value in the system. The state law gave both the administrators and the evaluators a chance to increase their authority.

The other district, 35, had had a management-by-objectives system for several years prior to our visits. It was quite routinized and was an important part of the district. It consisted of regular reviews of each teacher's and each school principal's performance and of each school's program. Reviews of school programs were done by teams, whereas reviews of teachers' and principals' performance were done in individual supervisor-subordinate dyads.

The evaluators' responsibility was to provide the appropriate evidence on a host of indicators -- things like test scores, attendance rates, or vandalism rates -- to each dyad or review team. Evaluators also kept records of all objectives that emanated from these reviews and provided progress reports to appropriate people at regular intervals. The evaluators in this district were very interested in the success of the management system, and adopted the management's point of view when they discussed it, saying, for instance, things like this:

> At the start of this system, we accepted very simple objectives, just in order to get the system accepted. We never said, "That objective isn't strong enough." But now we're starting to add tougher things.

On several occasions during our visits to this district, evaluators used the word "we" when explaining the rationale for the management system or when explaining that "we" are now tightening up the system, and they referred to teachers and principals as "they". On one occasion, one of the evaluators even switched to "I" when he said,

> Now [the system] may be scaring people. Now the goals are more sharply defined and we have changed our wording. In the technical memo, we say, " . . . ". This is a change in wording. It refers to observable events and milestones. That's what *I'm* really looking for. They are kind of like behavioral objectives [emphasis added].

Evaluators who adopt the role of management facilitator, then, tend to adopt the point of view of school district administrators. In that respect they are like their participant colleagues. But they are different from those colleagues in two other respects. First, participants tend to join their clients in brainstorming and in searching for patterns in the data, whereas management facilitators tend to provide the data and leave its interpretation to supervisor-subordinate dyads. Second, participants are as likely to rely on informal observations as on formal evidence when they analyze their clients' situations. Management facilitators, on the other hand, tend to confer their data with far more authority and to believe that it should be the primary stimulus in deliberations. The role of management facilitator is especially well suited to those organizations whose problem-solving strategy is hierarchically organized and procedes through cycles of goal setting and performance reviews. In school districts which rely on these strategies it may not be possible for the evaluator to adopt any other point of view, for these management systems place subordinates on the defensive, and the first line of defense tends to be that the data are not valid. Such a point of view would not be tenable for an in-house evaluator who hopes to maintain his or her position in the district over time.

The Independent Observers

The role of independent observer is closer to the ideal evaluator assumed in most evaluation literature than either of the other two roles are. Independent observers tend to be more concerned with providing technically credible or definitive information than with providing survey data or indicators that need judgment to be interpreted. Although they tend to serve senior administrators more than other client groups, they do not identify with any particular client group. Only two of the 16 evaluators participating in this study, those in districts 57 and 115, adopted the role of independent observer, and they did so in very different ways and for very different reasons. District 57's evaluation office had existed for over a decade, and its director had, over the course of that time, been able to establish and maintain a place for this role within the organization. District 115's evaluation unit, on the other hand, was relatively new. The problem-solving style in district 115 was highly political, and the new evaluators were anxious to demonstrate both their capability to produce accurate and timely information and their neutrality on the issues facing their clients.

District 57's evaluation office had established a number of procedural rules designed to assure objectivity. For instance, the director routinely rotated the evaluation staff among programs, so that no evaluator would become too involved with either programs or the people who ran them. In addition, the unit would not study topics that the director thought could not be objectively assessed. It would not study difficult-to-measure variables such as students' self esteem, it would not study difficult-to-document processes such as implementation, and teacher opinions about programs were considered out of bounds. The unit focused its attention on the effectiveness of programs and practices, and it measured effectiveness by means of standardized norm-referenced achievement tests. The director of the unit justified this emphasis as follows:

> I believe in achievement test scores. . . . [They] have their problems but they come pretty close to measuring what an individual child should be learning. It's not that I'm not interested in self esteem, but as far as I'm concerned, the most important outcome is basic skills.

The role this evaluation unit adopted had a predictable effect on the regard that members of the district had for the unit. Board members and senior administrators respected the unit and valued its contributions. School building staff, on the other hand, made several caustic remarks, claiming the unit did not know anything about instruction, that it caused more work and trouble than the help it provided warranted, that the department head was analogous to a high priest in a primitive society, and so forth. One teacher even said,

> They're over there with their computers and they don't always know how findings will affect schools. They think they just throw things out; they don't realize that the information they give has positive and negative effects and is used.

These reactions, both positive and negative, are the sort of reactions evaluators generally expect to receive when they adopt the role of independent observer. The client who must make the difficult budgetary decisions values the information, while those participants who are part of the ongoing programs tend to feel that their programs have been misrepresented or evaluated against the wrong criteria.

The role of independent observer was maintained not only because it was a role the evaluators believed in, but also because it was a role the organization had come to expect. Many of the methodological decisions that guided this unit were motivated as much by the need to uphold the appearance of irrefutability as by the need to be objective alone. For instance, one reason teacher opinions were not documented was that these opinions were not considered relevant to the objective worth of the program, but another reason was that these data could be challenged in a political forum. If the evaluators claimed that teachers held one opinion, those teachers who did not fit the norm could stand up and claim that the data did not reflect their views. In so doing, they could cast doubt on the validity of the entire study. Furthermore, the unit meticulously avoided any discussion of the possibility that achievement test data might be culturally biased, on the grounds that "It is best to let sleeping dogs lie." Finally, these evaluators were rarely able to randomly assign units to program options when they conducted summative evaluations, yet their evaluation reports did not discuss rival hypotheses regarding observed differences among groups. Many of the unit's reports had extremely large and detailed technical appendices, but these appendices described only measurement scales. Nowhere did the reports state where and how comparison groups were formed, or what implications the choice of comparison groups might have had for the kinds of inferences that could or should be drawn from observed differences. Rather than jeopardizing its credibility by openly discussing problems inherent in its data, this evaluation unit did not discuss either the appropriateness of its choice of measures or the appropriateness of its choice of comparison groups. Maintaining the image of the independent observer meant withholding pertinent information.

District 115's evaluation office, on the other hand, was new and was struggling to simultaneously develop a theory of evaluation and to insert itself into an organization that formerly had no need for evaluation. District 115 had the most political of all the problem-solving strategies observed during the course of this study. Nearly every issue went to the school board and was covered by local newspapers. The school board meetings were the forum for partisan debates. Its meetings routinely included testimony from parents and citizens with an interest in school issues, and the evidence produced by the evaluation unit were fed into these debates. This school district's organizational style had two important effects on the evaluators.

First, members of the district frequently spoke of winning and losing debates. This emphasis naturally made the evaluators want to win. On one occasion, they went to the board with a proposal to alter the district's testing policies. The district had been administering both achievement and ability tests to its students, and the evaluators argued that the data from these two tests were redundant and that there were important political and social reasons for not using the ability test. The board was skeptical and asked what other evaluation experts thought of this issue. Rather than maintaining its posture as advocate for change, the evaluation unit then reverted back to its posture of neutrality, and provided the board with a carefully orchestrated split panel of experts. Given this ambivalent expert testimony, the board decided to retain the tests. Members of the school district who referred to this decision tended to say that the director of the evaluation unit had "lost on that one." On another occasion, the evaluators entered into a dispute with one of the district's program directors over what questions the program's evaluation should address. In their zeal to develop a viable evaluation unit, they had developed a theory of evaluation that suggested the program director was asking the wrong questions. The issue went to the superintendent's cabinet for resolution. On that issue, the evaluation unit won and the program director lost.

Second, the highly political nature of decision making in this district made the evaluators more aware of the need to be neutral than of the need to produce evidence that was definitive by virtue of its technical virtuosity. In that sense these evaluators interpreted the role of the independent observer differently than did the evaluators in district 57. For one hotly contested issue, the school board specifically requested that a study be done by an outside contractor, so that it would be objective. The evaluation unit, anxious to demonstrate that it could handle such assignments, hired a contractor but worked closely with the contractor's staff throughout the conduct of the study. The cover of the report indicated that it was jointly authored by the in-house evaluation unit and the contractor, but members of the in-house unit were listed as first authors. The findings of this study were so neutral as to be almost useless. Every finding listed in the executive summary was an "On-the-one-hand/on-the-other-hand" statement. The evaluators made no attempt to weigh the contrary evidence they found or to use their own professional judgment to estimate what the bottom line on the program really was, for such an effort might have jeopardized their neutrality.

The two school districts in this sample whose evaluators had adopted the role of independent observer both had organizationally distinct units whose staff were labeled evaluators. But these evaluators did not assume their roles automatically, merely because they were evaluators. They faced two very important challenges. First, they had to define the role of the independent observer, and second they had to create a place in organizational decision making for that role. These two tasks were interrelated, for the creation of an organizational role depended in large part on creating and maintaining an image of credibility, and that image depended in turn on their definition of their role. For one unit, the role of independent observer was associated almost exclusively with technically objective measurement. That definition meant that many educationally relevant variables of interest to their clients were not studied on grounds that they could not be objectively measured, and it meant that the unit could not inform its clients of inferential weaknesses inherent in many of its evaluation designs. For the other unit, the role of independent observer was associated with political neutrality, and that definition meant that the evaluation unit could not use its professional judgment to sort out evidence for clients, but instead had to provide inconclusive reports.

SUMMARY AND CONCLUSIONS

Much of the literature on evaluation assumes that evaluators can define their own roles and that the greatest challenge facing evaluators is that of defining a role that consists of helping clients solve problems while simultaneously remaining independent of those problems. The thesis presented here is that evaluators' roles are determined to a large extent by the organizations they serve. In order to help their school districts solve problems, the in-house evaluators participating in this study adapted to the particular problem-solving strategies of their districts. They became technicians, participants, management facilitators, or independent observers. In those cases where the district experienced stress and changed its problem-solving strategies, evaluators were unable to change their roles accordingly and consequently their positions were jeopardized. In those districts where evaluators were able to adapt, their roles were compatible with their organization's needs, but not with the ideal evaluator role of helping clients solve problems while simultaneously remaining independent of those problems. Technicians produced data but gave no interpretive or other guidance to help their districts use the data. Participants adopted the perspective and the interests of their particular clients, often interpreting the data only from their clients' vantage point. Management facilitators and independent observers were often forced to sacrifice real credibility in order to preserve their image of credibility.

Successful adaptation had three effects. It assured continuing organizational support for the evaluation enterprise, it increased the practical value of evaluation products and services, and it meant failing to meet the professional standards for the evaluator's role. Yet none of these evaluators perceived their organizational contexts as compromising their professional obligations. Rather, the context merely reflected the clients' needs and in so doing defined the evaluator's job. The evaluators merely provided the services their districts needed. In their eyes, adaptation was not failure but success, and the unhappiest evaluators were those who could not adapt, for they could not serve.

NOTES

1. When units with labels other than "evaluation" perform such activities, I consider them to be in-house evaluators.

2. This selective attention to district-wide publications could also be because it is extremely difficult to determine what printout characteristics teachers and principals really find useful. In this study, for instance, nearly every district yielded within-district contractions among teachers regarding what was good and bad about their printouts and about what they wished they could have had.

3. District code numbers indicate the size of the districts in thousands of students served. District 4 serves 4000 students, and 240 serves 240,000. The code numbers randomly vary from real enrollments by ± 15%.

4. In this section I describe only the roles of the individuals or the units which constituted the primary evaluation activity in each district. If a district received information from both evaluation and curriculum units, I discuss only the activities of the evaluation unit.

REFERENCES

Barkdoll, G. L. Type II evaluations: Consultation and consensus. Public Administration Review, 1980, March/April, 174 - 179.

Bennett, C. A. and A. A. Lumsdaine (Eds). Evaluation and experiment: Some critical issues in assessing social programs. New York: Academic Press, 1975.

Boruch, R. F. and H. W. Reicken (Eds). Experimental Testing of Public Policy Boulder, CO: Westview Press, 1975.

Cohen, D. K. Politics and research: Evaluation of social action programs in education. Review of Educational Research, 1970, 40, 213 - 238.

Cooley, W. W. The inevitable subjectivity of evaluators. Educational Evaluation and Policy Analysis, 1980, 3, 89 - 90.

Kean, M. H. Compromising positions: The objectivity of evaluators. Educational Evaluation and Policy Analysis, 1980, 3, 87 - 88.

Krathwohl, D. R. The evaluator as negotiations facilitator-fact finder. Educational Evaluation and Policy Analysis, 1980, 2, 25 - 34.

Lindblom, C. The policy-making process Englewood Cliffs, NJ: Prentice-Hall, 1980.

Weiss, C. H. Organizational constraints on evaluative research. New York: Columbia University, 1971.

Wise, R.I. The evaluator as educator. New Directions for Program Evaluation, 1980, 5, 11-18.

Zeigenfuss, J. T., Jr. and Lasky, D. I. Evaluation and organizational development. Evaluation Review, 1980, 4, 665 - 676.

APPENDIX A

SAMPLING AND DATA COLLECTION PROCEDURES

APPENDIX A

SAMPLING AND DATA COLLECTION PROCEDURES

SAMPLING

Our original sampling goal was to obtain a pool of around 60 candidate school districts from which we could select 18 diverse districts to visit. Without direct knowledge about the total population of school districts, we decided to choose our sample on the basis of nominations. We wanted candidates to be nominated by a variety of people and for a variety of reasons. The process by which the eventual analytic sample was reached involved several stages.

Stage 1. The first step was to obtain a pool of candidate school districts by selecting people who were qualified to nominate and by asking them for information about candidate districts.

For nominators, we wanted people who had had direct contact with school districts rather than those who might have heard second-hand about districts, and we wanted people who could bring a variety of perspectives to the study. Our strategy was to select nominators by their affiliation with organizations that had different kinds of relationships with school districts. Table 1 summarizes the types of organizations we focused on and the number of individual nominators in each category from whom we received nominations.

As for the rationales for nomination, we recorded the evidence that nominators provided for each district. For each district, we also tried to obtain information on:

- the nature of evaluation or testing activities in the district;
- the way in which the data appeared to be used;

TABLE 1

Site Nominations

Source of Nomination	Number of People Contacted	Number of Independent Nominations*
Project Staff	5	25
Other Huron Staff	4	10
Other Contractors	9	19
Technical Assistance Providers	13	47
Federal Agency Personnel	7	12
University Personnel	8	26
Educational Associations	8	34
Test Publishers	4	23
National Consortium on Testing	2	6
SEAs and LEAs	3	10
TOTAL	63	212

* Many sites were nominated more than once. The total number of school districts nominated was 111.

- the names of people within the district from whom we could learn more; and
- the names of people outside the district from whom we could learn more.

Table 2 indicates, in abbreviated form, the types of reasons nominators gave for proposing that various school districts be visited. The table indicates that the majority of nominations were based on knowledge of the characteristics nominators assumed to be related to the use of evidence, rather than to indicators of evidence use per se.

Stage 2. Given the resulting list of school districts, we then called state education agency officials first to obtain support for or arguments against the nominations, and second to obtain other nominations. The reactions of state personnel were various. They included information to the effect that, for instance, a district had had an active evaluation office until the current year, when it was eliminated from the district budget, or that we had obtained only districts whose evaluators had Ph.D.s and had missed several smaller but very active school districts, or that the sample we had derived covered the best districts in the state.

Stage 3. By the time state agency staff had been called, we had acquired enough information about the districts that we could discriminate among them. The next step was therefore an attempt to reduce the number of candidates. Three independent reviewers read the entire data base and sorted the candidates according to their perception of the value of including the district in the study. Each reviewer was free to use his or her own criteria but all were agreed that there should be some indication that evidence did in fact tend to be used in the district and that there should be no indications

TABLE 2

Reason for Nomination

Reason Given	Number of Nominations
Nature of LEA: 59	
Superintendent	5
General attitude	14
Political orientation	6
Pending law suits	2
Solicitation of TA	16
Recent change in evaluation or testing	5
Evaluation Program or Director: 56	
Caliber of program	16
Caliber of director	8
AERA participation	8
User orientation	5
Staff involvement	2
Ties to decision-making process	3
AERA awards	12
Relationship with staff	2
Testing Program or Director: 4	
Caliber of program	0
Caliber of director	2
AERA participation	0
User orientation	2
Organizational Arrangement: 14	
Location of evaluation office	2
Multiple evaluation offices	3
Program vs. evaluation office	4
Contractual work	4
Other external evaluation	1
Methodology: 26	
Management Information System	10
Processes	1
Time-on-task studies	2
Computers	3
Rasch model; item banks	6
Standardized tests	1
NAEP or state assessment	2
Alternative assessment	1

(continued on next page)

TABLE 2 (CONTINUED)

Reason for Nomination

Reason Given	Number of Nominations
Uses or Purposes: 77	
Budget	1
Accountability	4
Policy Issues	6
Union negotiations	1
Needs assessment	3
Selection of Students for Programs	1
Title I	11
Other programs	8
Organizational or staff development	3
Counseling	1
Instructional reform	12
Mastery; competence	10
Diagnostic-prescriptive	8
Curriculum development	2
General emphasis or use	6
Bad Examples: 3	

that it had been misused. The raters then debated their ratings and developed a new reduced list of candidate school districts. About 40 school districts were retained for further consideration.

Stage 4. With a now much reduced set of districts to consider, we began calling members of the districts themselves, to learn what specific evaluation or testing activities occurred within the districts and to hear their point of view regarding how well these sources of evidence were used. We recorded their comments along with the others we had received already, thus increasing the data base on each candidate. We also asked for copies of evaluation reports, test printouts, or other materials that would indicate the nature of the district's evaluation and testing activities.

Stage 5. The materials we received from the districts, along with the additional comments they gave, provided the final addition to the data base. A number of tables were drawn up indicating how these remaining districts varied in size, geographic region, apparent major evaluation and testing activities, apparent primary audiences for these materials, and the nature of their state's policies regarding evaluation or testing. These summary charts were forwarded to the National Institute of Education and the U. S. Office of Education, and their staffs joined us in an iterative process of comparing trade-offs among alternative sampling plans until an initial sample of 18 school districts were chosen for visiting. One of these districts was replaced before our first round of visits, however, since the district chose not to participate in the study.

Stage 6. As field work progressed, the sample was further adjusted to accomodate initial findings. We began by paying a three-day visit to to each of these 18 school districts. Although the findings from all 18 districts were described in our first-year report, four of these districts were eliminated from this final analytic sample since for a variety of reasons, such as teacher strikes, these districts were not accessible for further data collection. Two other school districts were added during the following school year, bringing the total number of districts involved in this final analysis to 16.

SITE VISIT STRATEGY

Within each school district, one individual -- usually the evaluator -- served as our local host. These individuals assisted us in scheduling other interviews and making miscellaneous arrangements for the visits. They were also the first person we interviewed on arriving in the district and the last person we spoke with before finishing each visit. Interviews with these individuals gave us an opportunity to learn what kind of evidence the district produced and to whom it was given, and they provided useful overviews of the district and background information on other interviewees.

Given that orientation, we then tried to follow a sequence in which senior administrators were interviewed first, then program directors, building principals, and teachers. Our rationale was that each of these kinds of interviewees provided further context for succeeding interviews. The strategy was not always successful, of course, since the interviews had to be arranged at the interviewees' convenience rather than at our convenience. Within each visit's schedule, we also tried to hold some time periods open so that if need be we could re-schedule interviews with members who encountered emergency changes in their schedules, add new interviewees or additional interviews with former interviewees, or attend meetings we felt we might learn from. We also tried to schedule visits so that at least one school board meeting could be attended.

INTERVIEWING STRATEGY

The question of how people use evaulation or test data to change their strategies or improve their performance is a perplexing one, and a considerable body of literature is accumulating on the topic. Often information use is conceived of as a situation in which some kind of evidence is transmitted to an "audience," who then responds to the data. That is, the situation is viewed from the evaluator's perspective, rather than from the user's perspective. We chose an alternative point of view. We conceived of the information user as a person engaged in a set of problems related to his or her position in the district, and who draws upon information as it is needed.

We therefore designed our interviews to learn what our interviewees' jobs were like and what kinds of issues they faced. From this perspective, we could learn about how they tried to solve these problems, or resolve these issues, and how they used evidence in that context. This strategy meant both that our questions were very open-ended and that the topics we discussed with people varied considerable from person to person. Generally speaking, the sequence went like this:

Stage 1: Describe the Study.

Each interview began with a brief overview of the study, and we often gave people a brochure describing the study. This not only allowed us to introduce ourselves and our study, but also allowed interviewees to get comfortable with us before we began our interview. Several interviewees also needed to judge the value of the study before they were willing to spend their time with us. Sometimes this stage of the interview would last only a minute or two; at other times it would last as long as 15 minutes, as we responded to any questions interviewees had.

Stage 2.

After interviewees were comfortable both with us and with the study, we opened the interivews. We began with very broad questions about their jobs, either asking them what it was like, what they did, or what issues they faced. These questions got a wide variety of responses. Some refused to answer, preferring to go straight to the topic of how they used evaluation and testing information. Some read us their job descriptions and editorialized on each item. Some described a typical day. But most described their jobs more conceptually, telling us where they fit into the organization, what their responsibilities and goals were, or what the substantive issues were with which they dealt. From this point on, our questions followed the framework interviewees had provided for considering their positions. This stage of the interview often required as long as 45 minutes to complete, since the issues were complex and many had long histories that needed to be understood.

Stage 3.

Once we had a sense for the interviewees' points of view and things they were concerned about, we moved the conversation toward how they tried to resolve these issues. If possible, we would direct them toward an issue that was more likely to be informed by evaluation or test data. For example, if a teacher discussed both morale and the problem of identifying children with special needs, we would ask how the latter issue was resolved. Such strategies for directing the interviewees could not be applied uniformly, however, for even though an interviewee might mention three or four issues, his or her ensuing elaboration might make it clear that one of them was far more significant than the others. In these cases, the issues interviewees emphasized were pursued. As we found in other stages of the interview, people would discuss their methods of resolving issues in different ways. Some would describe their procedural methods -- prepare a statement of the problem, review it with the next level up, form a committee to make recommendations, and so on. Others would describe their substantive thinking, listing the several components of the issue, how they were related, how they would have to be balanced eventually, and so on.

Stage 4.

Discusion of how interviewees tried to resolve their problems would frequently open the door to questions about how data were or might have been used. If interviewees had volunteered a source of data, we would ask where it came from, whose idea it was to collect it, etc. If they quoted facts, we would ask how they knew those facts -- where the inforation came from. If no data were mentioned we asked if any data had been helpful, and if none were, what kind might have been. Sometimes, no convenient opening led to these questions, so that we had to be more resourceful. We could introduce a new topic by saying something like, "You've run this program for 10 years. Do you think it's improved over that time?" Or we might remind them of a report we knew existed, and ask whether it had been useful to them. Since stage 3 of the interviews often required the majority of the hour, there were cases where stage 4 was relatively short. These situations were rare, however, and usually occurred when the context provided during stage 3 was such that no elaboration was needed regarding the use of information, and a one-hour interview was sufficient. More frequently, our interviewees extended their time with us far beyond what had been scheduled. Administrators cancelled other appointments, and teachers took us with them to their classrooms after their free periods, gave children individual seat work and continued to talk to us. Our interviews with administrators often lasted 1-1/2 to 2 hours, and with teachers anywhere from 15 minutes to an hour.

APPENDIX B

PROFILES OF THE 16 SCHOOL DISTRICTS INVOLVED IN THE FINAL ANALYSES

Table 1
Sample Distribution by Geographic Location

Region	Number of Districts
West Coast	2
Southwest	6
Southeast	1
Midwest/Plains	1
Northeast	6

Table 2
Sample Distribution by Enrollment

Enrollment	Number of Districts
Less than 5,000	1
5000 to 15,999	2
16,000 to 29,999	5
30,000 to 59,999	3
60,000 to 199,999	3
200,000 or more	2

Table 3

Sample Distribution by Nature of State Mandates

State Policy	Number of Districts
Mandates Assessment	4
Mandates Accountability Process	4
Encourages Evaluation or Testing	2
No policies relevant to Evidence	6

Table 4

Sample Distribution by Organization of Evaluation and Testing Activities

Organizational Arrangement	Number of Districts
One Evaluation and Testing Office	8
Multiple Evaluation or Testing Units	4
No Formal Evaluation or Testing Unit	4

SITE 4

Five year enrollment change: Down 16%

Enrollment composition: Anglo 89%
Black 4%
Hispanic 1%

Number of superintendents over the past five years: One

Average per pupil expenditure: Unknown
Federal contribution: Unknown

Site 4 is a hold-over from the 1960's, primarily driven by its long-time superintendent, a charismatic leader with a strong Deweyan educational philosophy. The atmosphere is one of creativity and respect for other people's points of view. The central administration maintains a non-directive posture toward buildings and encourages building autonomy. Staff are encouraged to try new ideas, but the process is not very orderly and there is often little follow-though on new ideas.

The district is under considerable pressure to change now, both from the state, which requires a state assessment and a centrally-coordinated district-wide planning process, and from the community, which has taken on a "back-to-basics" attitude. The superintendent resists these pressures mostly by his personality and by pretending they are not serious threats. These tensions have, however, created dissention within the district staff.

The district has no evaluation office, but does have an "instructional resource" person who coordinates state and local testing and works with staff to interpret test scores as well as assist in curriculum developments. The district-wide view of tests is that they are not the best way to know a child. Other kinds of data available, such as enrollment data and the data produced by the accreditation process, are ignored or claimed to be of no use.

Site 4 was nominated because it engaged in an elaborate process of involving the community in its state-mandated planning process and produced interesting annual reports for the state and community. Site visits indicated that, although the process was lengthy and iterative, it was not systematic in its involvement of citizens, but instead relied upon those few parents who chose to attend occasional meetings.

SITE 7

Five-year enrollment change:	Down 14%
Enrolment composition:	Anglo 85% Black 13% Hispanic 1%

Number of superintendents over the past five years: Two

Average per-pupil expenditure:	$1466
Federal contribution:	8%

Site 7's administrative staff consists almost entirely of people who have moved up through the ranks, and many of them were natives of the town even before being employed by the district. In contrast to Site 4, the administrative style here is one of strict business and sound management.

The community served by Site 7 has changed considerably over the past few years. Middle income white-collar parents with college-bound children have been replaced by senior citizens and lower-income blue-collar parents whose children do not plan to attend college. The school district faces both declining enrollment and a declining tax base.

The administrators are responding to these changes systematically and aggressively. First, they engage in extensive public relations activities, especially when tax millage increases are to be voted upon. Second, they seek grant support from a variety of sources, and were administering 23 separately funded programs during the year of our visit. They have also obtained national validation for five of these programs, and have sought dissemination funds for these. Third, they have filled empty classrooms, both by offering preschool services and by offering a variety of adult education programs. Fourth, they are working to keep students from dropping out of school, in part by expanding their vocational and career education courses in secondary schools. They have also converted part of one of their buildings to a community center. Finally, they are trying to improve teacher evaluations in the hope that teachers can be non-renewed for cause, rather than being laid off according to seniority.

The district has no real evaluation office. It has a test coordinator who interprets and disseminates test results, performs a number of telephone surveys of neighboring districts to ascertain current practices in various areas, and calculates the statistics required for Title I evaluations. In addition, the high school guidance office conducts surveys of graduates and drop-outs, and consultants have been called in occasionally to assist in the preparation of evaluations for state or federal audiences.

Site 7 was nominated because of its large number of validated projects. Site visits suggest that people are proud of these projects, but were equally proud of other non-validated projects. Their motivation for seeking validation was not knowledge that these projects were especially good, but rather a desire to obtain dissemination funds. We found no evidence that evaluations played a role in decisions about these programs, though other kinds of data greatly influenced management decisions.

SITE 9

Five-year enrollment change: Down 17%

Enrollment composition: Anglo
Black (Unknown)
Hispanic

Number of superintendents over the past five years: One

Average per-pupil expenditure: Unknown
Federal contribution: Unknown

The most notable characteristic of Site 9 is its affluence. It is in a well-to-do sprawling suburban community whose parents expect a great deal of their children and of the schools. The nature of the community is such that neither funds nor student achievement present problems.

In response to the community, however, the district engages in a lengthy, iterative goal setting procedure in which goals are annually set at the classroom, building, and district-wide level, with the process repeated so that each level can accomodate the goals of other levels. The process is not only public but involves parents. It appears, however, to be more symbolic than real, since the attainment of goals is never addressed.

The district has an evaluation office consisting of two people who conduct mandated evaluations, administer testing programs, and review research literature relevant to current issues in the district. Only the literature reviews were referred to by participants in this study.

Site 9 was nominated because it was trying to develop a state-of-the-art preschool screening program and had hired expensive consultants to help. Site visits suggested that their major concerns were with developing the preschool program itself more than with developing early childhood assessment techniques.

SITE 17

Five-year enrollment change: Down 28%

Enrollment composition: Anglo
Black (Unknown)
Hispanic

Number of superintendents over the past five years: Unknown

Average per-pupil expenditure: $1500
Federal contribution: Unknown

Site 17 is a suburban school district, established in the early 1940s by consolidating several smaller suburban districts. Its enrollment rose dramatically in the 1960s and declined just as dramatically in the 1970s, to about half its largest size. During its first three decades, the district encouraged variation among its schools, at first because independence was palatable to the several constituents in the new district, and later because enrollment increases absorbed the attention of the administration. Throughout this period, the district viewed diversity as a strength and maintained an open enrollment policy in its elementary schools. But in the 1970s the district not only lost a significant part of its enrollment, but also lost two important tax millage elections.

In response to its perception that the public had lost confidence in the schools, and in response to a new state law, the administration established a uniform set of learning objectives and a computerized individualized testing system. It is using the state law to justify mandating this system.

The evaluation unit in the district is responsible for the computerized testing system, standardized testing, enrollment projections, and mandated evaluations. It also conducts a number of ad-hoc analyses of test data to assist administrators in planning and oversight. The staff are more appreciated and more highly revered by administrators than by teaching staff, who are still leary of the computerized assessment system.

Site 17 was nominated because of its computerized testing system. Site visits indicated that teachers could test any combination of students on any combination of objectives at any time. Not all teachers like or use the system, however. Further, teachers still trust their own judgment more than the computer's, and the system is designed to allow them to override the computer's judgment regarding whether or not a student has mastered an objective.

SITE 18

Five-year enrollment change:	Down 12%	
Enrollment composition:	Anglo	42%
	Black	46%
	Hispanic	12%

Number of superintendents over the past five years: Three

Average per-pupil expenditure: Unknown
Federal contribution: Unknown

Though relatively small, Site 18 suffers from big-city problems: an increasingly minority population which is segregated from the white population, declining enrollment, high student absenteeism, transience and misconduct, a declining tax base, and a relatively less educated, conservative community with high unemployment due to faltering local industries.

The staff within the district is divided on almost every issue--emotions run high and there are a number of political factions and hidden agendas. The teachers' union is strong and engaged in a lengthy strike prior to agreeing on its most recent contract. Most aspects of the district appear to be in dissarray, and decisions are usually made by a small group of people who keep their cards close to the vest. The atmosphere is one of cynicism and distrust.

The district has an evaluation unit with seven members which administer the testing programs, conduct mandated evaluations, manage an extensive information system, and engage in a variety of small, special-purpose studies, such as literature reviews and telephone surveys. With the exception of the superintendent's reliance on management information, there was little evidence that much of this information was used. Given the atmosphere, however, it was hard to determine the real basis for most points of view and decisions, though several people spoke of suppressing information so that it would not get into the "wrong hands".

Site 18 was nominated because it had an outstanding evaluation unit which prescribed to the CIPP model of evaluation. Site visits revealed, however, that the evaluators themselves claim the model is not politically feasible to use.

SITE 19

Five-year enrollment change: Up 29%

Estimated enrollment composition:		
	Anglo	90%
	Black	2%
	Hispanic	8%

Number of superintendents over the past five years: One

Annual per-pupil expenditure: Unknown
Federal contribution: Unknown

Site 19 is a secondary school district (spanning grades 9 - 12) and resides in a predominently white working class suburb. It has had stable management for over a decade. The school district is growing and has not faced major policy issues for some time. Consequently, it has been able to concentrate more than most districts on its curriculum. For the past seven years, it has been developing an instructional management system which consists of instructional objectives, criterion-referenced tests, and teacher training activities for all required courses in the curriculum.

This system has affected decision making at two levels. When the board is faced with budget cuts, it tends to cut courses that are not part of the instructional management system. At the classroom level, teachers gear their efforts exclusively toward tested course content. The central administration, on the other hand, administers the system but appears not to be affected by the data it generates.

District 19 has no evaluation office. Its massive testing program is administered by the curriculum department, and it has received occasional assistance from professional evaluators in other school districts.

District 19 was nominated because of its use of tests for instructional management. Site visits indicated, however, that teachers cheated in several ways so that their classroom test score averages would appear to be acceptable.

SITE 25

Five-year enrollment change: Up 21%

Enrollment composition:	Anglo	12%
	Black	5%
	Hispanic	83%

Number of superintendents over the past five years: Two

Average per-pupil expenditure: $1400
Federal contribution: 17%

The biggest problem facing the Site 25 administration is the district's rapidly growing population. The school district receives both legal and illegal immigrants from Mexico at such a rate that the district's budget is constantly geared toward building new facilities. As a result, the administration is quite small, relative to the size of the student body, and, with the exception of one building which has federally-funded computer-assisted instruction, there are no instructional frills.

In addition to rapid growth in its student population, this district faces a student population that ebbs and flows, since many of the students are from migrant families. The district works a lot with parents to convince them of the benefit of postponing their migration until the end of the school year. Site 25 was the only school district we visited that included the community in its organization chart.

Since the superintendent's primary concerns are with physical plants, he has little use for research and evaluation. Until this year, the district supported a one-man evaluation unit to attend to federally-required evaluations. In an effort to streamline these several requirements, the district instituted a district-wide norm-referenced testing program four years ago, and administrators are now beginning to use these test results to monitor instruction.

Site 25 was nominated because it was a district with a previously untrained evaluator who had profited from Title I Technical Assistance in evaluation and had conducted strong evaluations of the district's Title I programs. Site visits revealed that this individual had extended his training to the evaluation of other categorical programs as well, but also that his position as evaluator is being abolished. He would become a program director the year following our visit, leaving the Title I director to conduct his own evaluations in the future.

SITE 27

Five-year enrollment change: Up 10%

Estimated enrollment composition: Anglo 90%
Black 2%
Hispanic 8%

Number of superintendents over the past five years: Two

Average per-pupil expenditure: $1500
Federal contribution: Unknown

Site 27 is an elementary school district (spanning grades Kindergarten through eight) and serves the same community as Site 19. It is predominantly white, conservative, and working class. Site 27, however, recently acquired a new superintendent who had new management ideas. He introduced management by objectives, planning cycles, and criterion-referenced tests to the district. He also introduced merit pay for administrators and principals. Teachers and pricipals are evaluated in part on the basis of their student's test scores.

These innovations have introduced some stress into the system. Several staff members felt the changes had been made too quickly, and teachers added that there was too much testing and too much pressure to teach to the test. Though teachers were not part of the merit pay system, their principals were, and so teachers felt the pressure.

Site 27 has no local evaluation office, but has a senior administrator in charge of policy, planning, and evaluation. It has received considerable assistance from a neighboring school district and from the local university.

Site 27 was nominated because it was an example of a district which had no evaluation office of its own, but which had profited from methodological assistance from outside sources. Site visits confirmed this observation.

SITE 35

Five-year enrollment change: Up 25%

Enrollment composition: Anglo 86%
Black 1%
Hispanic 10%

Number of superintendents over the past five years: One

Average per-pupil expenditure: $1517
Federal contribution: Unknown

Site 35 resides in a predominently white community that reflects a stereotype of "middle America". The staff tend to have the personalities of cheerleaders and the district as a whole seems to view the local business community, rather than parents per se, as its most important constituency.

Perhaps because of its business influences, the district places heavy emphasis on performance accountability. Nearly every aspect of the district is managed by objectives, and the district uses a variety of planning and goal-setting procedures, all of which are compatible and appear to be well coordinated. Unlike other districts which rely on such systems, staff in Site 35 were not afraid of these procedures and only rarely seemed to try to undermine them.

The district supports a formal evaluation office of five people. In addition to managing the testing program, this office maintains an extensive management information system with data pertinent to most goals, conducts all the mandated evaluations, and occasionally conducts special management studies. Most of its studies consist of surveys of attitudes toward and perceptions of the educational process.

Site 35 was nominated because the evaluation office provided studies designed specifically for policy development. Site visits indicated that such studies were done occassionally, but that they did not comprise a significant portion of the evaluation work-load nor a significant portion of the data that policy-makers found to be useful.

SITE 50

Five-year enrollment change: Down 25%

Enrollment composition: Anglo 60%
Black 20%
Hispanic 4%

Number of superintendents over the past five years: Two

Average per-pupil expenditure: $2745
Federal contribution: Unknown

Site 50 faces two difficult problems. First, its environment is rapidly changing. Not only is overall enrollment declining but minority enrollment, particularly non-English-speaking, is increasing. The district is finding it difficult to serve its nearly 70 different language groups of students while desegrating the schools and receiving a smaller and smaller financial appropriation from the state. Its second problem is internal. A new superintendent has reorganized the central office and decentralized several responsibilities in order to remove power from senior administrators.

These problems have left several decision-making voids, and many of our interviewees discussed the difficulties of getting their work done or the difficulties of getting decisions made, more than the substance of what the issues themselves were. Everyone was preoccupied with coordinating with one another and with complying with desegregation and the Lau decision.

Site 50 supports a small evaluation unit which is responsible for all mandated evaluations, the testing programs, and some ad-hoc studies needed for complex administrative decisions. Participants in this study referred more often to the personal assistance of the evaluators in planning for desegregation than to any formal documents or information produced by the unit.

Site 50 was nominated because the evaluation unit had won an award for a study done several years ago, and because of the way it was implementing a new state assessment law. Site visits provided no elaboration on these activities, however, since neither was sufficiently relevant that they were discussed by interviewees.

SITE 57

Five-year enrollment change:	Down	1/2 %
Enrollment composition:	Anglo	58%
	Black	18%
	Hispanic	29%

Number of superintendents over the past five years: Two

Average per-pupil expenditure: $1755
Federal contribution: 8%

The major problems facing Site 57 had to do with coordinating multiple categorical programs and preparing for a forthcoming desegregation effort. These issues were by no means overwhelming, however, and most district staff at both central and building level focused on instructional and programmatic improvement issues. There was little evidence of power struggles or other disruptive influences on practice.

Site 57 has recently begun to take test scores very seriously. The state has initiated a graduation competency test and is considering an assessment as well. The school board has begun to systematically discuss test scores and to inform building staff that they are pleased or displeased with their performance. So far these efforts take the form of pressure in the air more than concrete policies, but they are influencing teacher and principal behavior.

Site 57 maintains a ten-person evaluation office which administers the district's testing program, conducts all mandated evaluations, and conducts a variety of special purpose studies. The special purpose studies appeared to be unusually influential, both in the central administration and in the buildings, though they influenced thought more than action.

Site 57 was nominated both because the evaluation unit was responsive to board concerns and because it made a point of identifying client information needs prior to conducting studies. Site visits indicated that the board was not very aware of the evaluation unit or what it did, but that the unit usually did work closely with clients to formulate researchable questions before beginning new studies.

SITE 72

Five-year enrollment change: Down 22%

Enrollment composition:	Anglo	42%
	Black	44%
	Hispanic	11%

Number of superintendents over the past five years: Three

Average per-pupil expenditure: $2200
Federal contribution: 12%

Site 72 was the most chaotic district participating in this study. Decision-making often consisted of participants accusing one another of hidden agendas, and parents, the courts, and the city government routinely interviewed in district activities, though often not very constructively. These battles also meant that no one knew what their budget would be more than a few weeks before the beginning of the school year, and budgets often changed capriciously during the year.

Distrust is particularly high between central administration and the school buildings, and this often centers on the accuracy of available management data.

The district supports a variety of evaluation units, none of which has remained stable for sufficient time to develop working procedures or a set of interested clients. Most of the data produced were descriptive statistics and we found very little use even of these.

Site 72 was nominated because the evaluation unit had won an award and because of the quality of its management information system. Site visits indicated, however, that the evaluator who had won the award was no longer an evaluator, and that district staff argued among themselves continuously about the quality of the information system. One nominator, however, also indicated that decisions were mostly political, but site visits suggested that they were too capricious to have even been politically motivated.

SITE 83

Five-year enrollment change: Down 1%

Enrollment composition:	Anglo	50%
	Black	3%
	Hispanic	38%

Number of superintendents over the past five years: Four

Average per-pupil expenditure: $1870
Federal contribution: 11%

Site 83 is in the middle of a painful transition. The old style of management was decentralized and variegated. Individual principals could lobby for funds for programs of their own making, or for programs they had seen in other districts. The district encouraged visits to other districts. Evaluators were distributed throughout the district to help principals evaluate their efforts, but negative summative evaluations rarely led to curtailment of funds from the central office. The new style of management, not yet in place, will be one of predominently uniform curriculum, and decisions regarding the funding of special programs will be centralized and more probably based on evaluation data.

The transition has been difficult. There have been three superintendents in the past two years, and several other members of the district have either gained or lost considerable power as their programs have been centralized, revised, or deleted. The administrative turnovers, coupled with rapid changes in organization, have produced both tensions and ambiguities. Many people cling desparately to friendships and allegiances to preserve their territory, and decisions tend to be based heavily on personal alliances, personalities, or friendships.

The evaluation function is one of the newly centralized activities, though individual evaluators still serve the same clients and provide essentially the same services. The primary evaluative activities still consist of administering the testing program, conducting mandated evaluations, and helping building principals and teachers evaluate their own activities and interpret test scores. Many of the studies designed to assist central management are contracted out.

Site 83 was nominated because the Title I evaluator was supportive and helpful to building staff. Site visits indicated that teachers appreciated this evaluator's services, which consisted primarily of item analyses and consultations. Non-Title I teachers had similar regard for their own evaluator-consultants, but were envious of the Title I teachers' access to item analyses.

SITE 115

Five-year enrollment change: Down 16%

Enrollment composition: Anglo 80%
Black 11%
Hispanic 3%

Number of superintendents over the past five years: Two

Average per-pupil expenditure: $2499
Federal contribution: 7%

Site 115 resides in a large suburb of mostly upper-middle-class whites. The parents expect a great deal from the schools and routinely challenge school district decisions. The district has traditionally had high-scoring students, taught by highly educated and well paid teachers and highly educated parents. Because parents care about all aspects of schools, the school board does too, and tends to get involved in far more than budget issues.

The atmosphere in this district is one in which every issue is intensely scrutinized by everyone and every decision is preceded by heated, though mutually respectful, debate. Arguments are nearly always data-based, including those presented by parents. Everyone seems to know that if they want to participate in policy discussions, they must base their arguments on sound data and they must expect their data to be challenged.

The district supports several evaluation units: one specifically for enrollment statistics, one within the Title I program, and a third, much larger one, which administers the testing program, condusts a variety of special-topic studies for the school board, and monitors occasional studies that are contracted out to insure objectivity.

Site 115 was nominated because of the quality of its annual report on test scores, because it addressed a variety of management and policy needs, and because it conducted studies of special topics such as time on task and was attempting to relate time on task to test scores. Site visits indicated that the annual report on test scores was widely distributed among staff and the public, that policy-makers appreciated and attended to most of the studies that were done for them, and that no one ever heard of any studies of time on task.

SITE 220

Five-year enrollment change: Down 12%

Enrollment composition: Anglo 30%
Black 63%
Hispanic 7%

Number of superintendents over a five-year period: One

Average per-pupil expenditure: $2822
Federal contribution: 17%

Like many large urban school districts, this one is plagued with law suits involving use of funds, hiring practices, provision of services to handicapped youth, and desegregation. In addition, it has declining enrollments and an increasingly disproportionate enrollment of minority youngsters.

Despite these facts, however, the district seemed to hum along and to maintain a very bureaucratic and staid atmosphere. The central administrators were for the most part optimistic about their district and seemed to be working to improve student performance. Nearly everyone interviewed was aware of the pattern of test scores in the content area or regional area they were responsible for, and had hypotheses regarding the causes for high and low scores.

Site 220 supports a very large evaluation unit which administers the districts' testing program, conducts all mandated evaluations, and maintains a computerized information system. The data most frequently mentioned by interviewees in the district were the test data and the enrollment data, the former apparently greatly influencing programmatic thought and the latter influencing decisions regarding school closings and desegregation.

Site 220 was nominated because of the caliber of the evaluators employed in its evaluation unit, particularly the director. Site visits indicated that, while an enormous volume of evaluation reports were produced annually by this staff, the bulk of these reports followed a common reporting format that was dry, and there was very little evidence that these studies influenced any of the district decision makers.

SITE 240

Five-year enrollment change: Down 8%

Enrollment composition:

Anglo	40%
Black	29%
Hispanic	31%

Number of superintendents over the past five years: Three

Average per-pupil expenditure: $2000
Federal contribution: 13%

Like many large urban school districts, Site 240 faces declining enrollment, a disproportionate increase in minority enrollment, and budgetary problems. The central administrative offices have been reorganized several times in the past five years, and many central administrative offices have lost more than half of their staff. These changes, combined with rapid turnover of superintendents, have led to low morale.

The main concerns discussed by interviewees were the difficulties of maintaining adequate levels of services with reduced staffs, and the problem of coordinating efforts across such a large bureaucracy when resources are constrained and student needs are rapidly changing.

The evaluation activities in Site 240 are limited to three: required evaluations, a district-wide testing program (which includes a state assessment) and a system of monitoring building-level compliance with local, state, and federal rules. Lack of staff and funds prohibits the unit from going beyond rudimentary involvement even in these activities.

Site 240 was nominated for this study because it had a solid evaluation unit and had won an award for one or more of its studies. However, by the time of our visit, the unit had been reduced considerably by a district-wide reduction in force.

APPENDIX C:

THE ANALYSIS OF CLAIMS ABOUT THE USE OF EVIDENCE

THE ANALYSIS OF CLAIMS ABOUT THE USE OF EVIDENCE

Lack of knowledge about how social science evidence such as evaluation and test data are used has led many investigators to study the issue by means of open-ended, qualitative research methodologies. These methods are suitable either when very little is known about the topic under investigation or when the topic itself has to do with subtle or only partially-conscious processes. Both of these conditions obtain for the subject of how social science evidence is used. But open-ended investigative techniques do not automatically expose the topic is interest. For instance, the topic of how evidence is used presents several special problems, the most important of which is that the use of evidence is something that occurs within the user's head. It is not directly visable and it can occur at such odd moments as while the user is driving to work, taking a shower, and so on. Knowledge of how evidence becomes encorporated into the thoughts of users, or of how it actually changes those thoughts, may be as elusive to the user as it is to the investigator. Simply being present in the field, therefore, is not sufficient to expose the process. Because researchers cannot directly observe the phenomenon they want to study, they must instead ask users to describe it, and they must accept their users' verbal claims as their data.

And that fact leads to another problem, for interviewees may not be able to express very precisely the exact contribution that evidence has made to their thoughts and actions, even if they think they know and are willing to try to explain it. They may say that a study was "really helpful", that it helped them "think about things", or that it "helped with the budget." Though field investigators can ask further questions to clarify these claims, the responses they receive will always depend in large part on how introspective and how articulate the respondent happens to be.

Furthermore, there may be situations in which users are completely unaware that they have used evidence. Users may be aware that they are basing their ideas on evidence at first, but may later on perceive the same ideas as "common sense". Perhaps still later these ideas become the assumptions on which other ideas are based. When later decisions are made, users are unaware that they are based on evidence.

Finally, users may on occasion claim uses that are not really true. Rhetoric regarding how evidence should be used is not limited to the evaluation community alone, but is part of the belief system of practitioners as well. Those observers who try to learn how evidence is used by asking direct questions are likely to hear what practitioners think is the "right" answer, rather than hearing what practitioners really do. Such claims may not be made because users intend to falsify the record, but may be made because users believe they behave differently than they really do.

Because of these problems, researchers who enter the field to get a "naturalistic" look at this phenomenon often discover on arriving that the phenomenon still defies observation, and that they must rely instead on verbal claims about the phenomenon. Though such claims may be valid indicators of use, they are not direct evidence of use. The investigator needs rules for assessing the validity of these claims. The challenge in designing a field investigation is to develop methods that are sensitive to two kinds of errors: false positives, or cases in which practitioners claim to use evidence in ways which they really do not; and false negatives, or cases in which practitioners fail to claim that their ideas or actions were influenced by evidence when they really were.

In addition to assessing the validity of alternative kinds of claims, and eliminating invalid claims from the data base, the researcher must also rely on those claims that remain in the data base to define the phenomenon under study. Consequently, even if the investigator is meticuluous in eliminating all invalid claims, he or she is still at the mercy of the remaining claims. They are still only indicators of use, rather than direct evidence of use, yet they are the only data available. The investigator must develop rules for determining how these indicators can and should be interpreted, so that they can be considered to be not only valid but illuminating as well.

This paper describes a study of school district uses of evaluation, testing, and other sources of social science evidence, and illustrates both data collection and data analysis procecures that can be used to compile a body of valid and informative claims. The paper has three main sections. The first describes the study's sampling and data collection procedures, the second describes the rules used to eliminate irrelevant or invalid claims, and the third shows how the remaining claims were sorted and what they were taken to indicate about the use of evidence in school districts.

SAMPLING AND DATA COLLECTION

This investigation entailed visits to 16 school districts during the 1979 and 1980 school years. The districts were sampled from a pool of 120 candidate districts which were known to have used some form of evaluation or test data recently. The original 120 districts were nominated for a variety of reasons, including, for instance, that a district had recently desegregated its schools and had used enrollment data to do so, had relied on research to settle a teacher strike, had established a district-wide curriculum testing program, had responded creatively to a state-mandated self-assessment, had a large or vigorous or capable evaluation staff, or had a creative director of the testing program. Since the nominators themselves came from all educational walks of life and appeared to use a variety of different standards for nominating school districts, the implications of this sampling procedure are not known. However, because 16 of the districts were eventually visited, it is known that there was very little relationship between the reasons these 16 districts were originally nominated and the activities occurring in these districts at the time they were visited.

About half of these nominations were rejected either because knowledge about them was too skimpy or anecdotal or because they were nominated by only one individual and no seconds for the nomination could be found. Selection from the remaining half was an iterative and judgmental process in which an attempt was made to ensure that visited districts would vary in the following ways: presence/absence of an evaluation office, functions of the evaluation office if there was one, geographic location, size, wealth, and ethnic composition of the student body. The final sample included districts from fourteen states scattered across the contiguous United States. These districts ranged in size from serving 4,000 students to serving 240,000 students and in ethnic composition from mostly white to mostly black to mostly Hispanic. They had per pupil expenditures ranging from $1400 to $2700 and had federal contributions ranging from 4 to 17 percent. Four had no evaluation offices, three had multiple or decentralized evaluation units, and the remaining eight had centralized units which varied in the nature of services they provided and the audiences they tended most to serve.

Each district was visited for ten to sixteen person-days, and the data gathered during these visits were entirely qualitative, coming from interviews with individual members of the districts and observations of meetings held within the districts. Samples of the evaluations, tests, or other formal evidence referred to during these visits were also gathered. Observations within each district were limited to those meetings that happened to occur during the time of the visits, and to which field investigators were permitted access.[1] Interviews, on the other hand, were scheduled in advance and were designed to include members of the policy-making community (usually the superintendent, assistant superintendents, and school board members), the program development community (usually program directors, curruculum coordinators, and supervisors) school building principals and classroom teachers.

To avoid the possibility of eliciting false positives during private interviews, participants were asked to discuss issues that they themselves were concerned about, and how they were trying to resolve those issues, rather than being asked to discuss their use of evaluation or test data per se. Once evaluation data or test data were brought up in the context of such an issue, the interviewer would then ask more questions about where the data came from, how the respondent knew about the study, and so on, while still retaining the general tenor of the conversation as one about how the practitioner was resolving a substantive issue. The intent, then, was to elicit references to evaluation and test data only when they fit into the practitioner's natural train of thought, if they did at all. To avoid false negatives, respondents were frequently asked how they knew something, why they believed one thing rather than another, or why they predicted one outcome rather than another. These questions brought forth references to evidence that had been assumed as background knowledge by interviewees. Field investigators did not intervene during observations of meetings, of course, though they often sought out participants for private interviews later on.

Nearly 400 notes resulted from these interviews and observations, with a typical note being eight pages long and describing a one-and-a-half hour interview or a two-hour meeting. The substantive content of these notes ranged from issues facing classroom teachers to those facing principals, program managers, superintendents and board members. They included such things as whether to place a child in special education, how to find space for a bilingual program without upsetting desegregation, settling a teacher strike, choosing a school to close, deciding whether a policy needed revision, and assigning course grades to students. From this mass of notes, only those references to formal evidence were of interest. The notes were therefore coded, and all references to formal evidence were pulled from them. Rather than prejudging references according to whether they should be considered "legitimate" uses or not, the coding was designed to obtain as broad a coverage as possible.

ELIMINATING IRRELEVANT AND INVALID CLAIMS

Figure 1 illustrates how citations were derived from the notes. There were occasions, and Figure 1 demonstrates one of them, when a member of a school district might utter multiple references to evidence almost within the same breath. On such occasions, the references were coded as separate, very small citations. But there were also cases when much larger citations appeared. For instance, a participant might devote a paragraph to one source of evidence, and do so in such a way that this paragraph would constitute a single citation.

FIGURE 1: Illustration of codes drawn from text of field notes

---

Text of an observation of a school board meeting

. . . The superintendent described some modifications he wanted to
make in two school buildings. The first was the [ABC] school, which parents
wanted changed from a K-6 school to a K-8 school. The superintendent proposed
that in order to do this, the seventh and eight graders would attend the
[DEF] school for some of their classes, but would remain primarily at the
ABC school. By way of reassuring the board of this plan's merits, he said
(1) the ABC school had a good atmosphere, was well run, had high test scores,
(2) and was a racially integrated school. I found it interesting that he included
test scores in his description of the school since so many of the districts
we have visited are reluctant to consider test scores when they make these
kinds of changes. The superintendent then pointed out that this change
would be consistent with the district's reorganization and desegregation plan,
which [involved a number of other changes]./ He then turned to his assistant
superintendent and asked if he wanted to add anything to this discussion.
The assistant superintendent pointed out that the DEF school had moved from
(3) 5% white to 20% white. The main question asked by board members had to do
(4) with whether this change would rob the DEF school of the cream of its crop,
a reference to the fact that only the brighter students would be transfered.

Furthermore, once the coding was completed, certain clusters of citations were combined and labeled as single citations. This occurred when the investigator observed a meeting in which the meaning of a study was debated for some time, or in which an issue was debated and a particular study was referred to repeatedly. The original coding of such an event might yield several dozen citations, and artificially inflate the number of citations coming from a particular district. Yet if the observer had not been present, but had heard of this meeting from a participant, there might be only one or two citations from the interview note. Therefore these clusters were coded as single instances.

Pulling these citations from the notes was not only time-consuming but also extremely tedious, and consequently less attention was given to inter-coder reliability than would be desireable. Only a handful of notes were coded by more than one reviewer, and the inter-coder comparisions suggested that there was about a 10 percent difference in the total number taken from each note. These double codes were checked again after the data were sorted, however, and found not to differ in the relative frequency with which different types of citations tended to be identified.

All citations were pulled from their original notes and assigned code numbers indicating the district, the context (policy, program, building or classroom), the particular note, and the page number from which the citation was taken. There were 2,975 citations altogether, and these constituted a new data base which could be analyzed to determine how social science evidence was used in school districts. But before engaging in that analysis, false positives and other irrelevant or invalid claims had to be cleaned from the data base. Several kinds of citations were eventually deemed to be either irrelevant to the use of evidence or to be relatively less valid than other citations.

The first citations to be removed were those in which respondents mentioned some procedural detail regarding how data were collected, analyzed, or disseminated. These were deleted on the grounds that, although they suggest that respondents at least knew that data were available, the citations themselves do not indicate that the respondents actually did anything with the data. There were 514 statements describing details of data collection or dissemination procedures. They included such comments as these:

> - - District 115, Principal[2]:
> We prepare charts on expected performance level for each grade level.
>
> - - District 72, Title I Director:
> The MAT [Metropolitan Achievement Test] is used to conduct our Title I evaluations.

Second, since the original coding purposely cast as broad a net as possible, it was necessary to eliminate those citations that referred to informal information, such as personal observations or rumors, rather than formal evidence. This exclusion rule resulted in retaining citations that referred to management information such as enrollment, attendance, vandalism or drug abuse statistics, test data of all sorts, surveys, correlational studies, evaluations, and anything called a "study", an "evaluation", or a "report". Eliminated were 312 citations such as the following:

> - - District 18, secondary principal:
> I look for small indicators of change. For example, four years ago, this school was filled with racial

> tension. You rarely saw teachers and students informally talking to one another. The blacks sat in one part of the cafeteria and the whites in another. . . . These are the indicators that count.

- - District 220: Superintendent:
  Last year for the first time, the senior hight school won the swimming championship. That's evidence of the benefit of having a swimming pool in the junior high.

Third, despite the attempt to target interviews on substantive issues rather than on the use of evidence per se, several citations contained opinions regarding how evidence is or should be used. These comments indicate the prevalence of rhetoric regarding the use of evaluation and testing, and their substance would make an interesting study in itself. But they were eliminated because they do not indicate whether evidence is in fact used in the ways stated. 608 opinions regarding how evidence is or should be used were eliminated from the data base, and these include the following:

- - District 4, Principal:
  Being able to make comparisons on these [test scores] should be of use to me as a supervisor.
- - District 72, Administrative Assistant to the Superintendent:
  Hard data can tell you that your reading program isn't working. What to do about it depends on the soft data. I'm not sure if that's how it should be, but that's how it is. Educators don't use R&D.
- - District 83, Title I director:
  One of the major plusses of federal programs has been the concept of evaluation [and] needs assessment.

There were also some comments which were eliminated on the grounds that they were hearsay. They described what the speaker thought someone else had done with the evidence. 106 hearsay statements were eliminated, and the three hearsay claims shown below indicate why. All came from the same school district and all are hearsay claims about how teachers in that district use the district's norm-referenced achievement test.

> -- District 25, program director
> The teachers do not like the time it takes, but they do think it is something solid to show how they are doing.
>
> -- District 25, program director
> We try to make the teacher aware of the results, and we hope the teacher might learn something from them. But no matter what the results are, it does not affect the teacher.
>
> -- District 25, program director
> It varies from teacher to teacher. Some of them have a blind faith in testing, are hung up on testing and make unwarranted interpretations. On the other hand, there are also teachers who tend not to believe the tests so much and tend to see them as secondary to their own judgments.

Finally, there were some citations in which respondents claimed that a source of evidence had not been used. These were eliminated primarily because their meaning and consequently their relevance was not clear. Some verged on hearsay, stating that the district had not used a study; some verged on opinion, offering a reason why evidence was not used; and some were too brief to have any meaning. There were 126 references to occasions in which evidence was not used, and they included the following:

- - District 57, Program Director:
  I ignored this study. [The study suggested that program directors in the district provided too much positive feedback]
- - District 4, Superintendent
  I have no use for data collected for no use, especially mandated data.
- - District 35, Principal:
  [The district criterion-referenced test] is not useful. The turn-around time is too slow.

The removal of these five kinds of citations reduced the data base from 2,975 citations to 1309 citations, less than half the original set. However, a second review of the citations was inspired by the work of Becker, Geer, Hughes and Anselm (1964), who attempted a similar analytic strategy. These authors not only sorted citations by the substantive arguments they made, but also according to whether the citation was elicited by an interviewer's question versus being volunteered by the respondent, and whether the citation came from a private interview versus from a group setting. Their reasoning was that volunteered statements and statements made in group settings were more likely to represent true beliefs and behaviors. Their analytic strategy suggested the notion that claims could vary in their degree of validity, and suggested the possibility of a second round of eliminations from this data set. The elimination rules used during the second pass were not the same as those used by Becker and his colleagues for several reasons. First, there were occasions in this study when a claim might be elicited by an interviewer during a conversation about a substantive issue, so that it would not be clear that the interviewer's question really made the respondent more self-conscious about his or her use of evidence per se, even though the question may have sensitized the substantive issue. Second, elimination of all private interviews would have substantially reduced the size of this data base, and may have done so unnecessarily. Becker and his colleagues were interested in group attitudes and mores. For that topic, evidence gathered from group observations may have been far more valid than evidence taken from individual interviews. Since our topic dealt with private uses as well as public uses, we were less convinced that evidence from private interviews were invalid. Finally, Becker and his colleagues did not eliminate, but instead merely separated, their less valid claims from their more valid ones. Since this analysis entailed elimination, there was more interest in assuring that claims were really invalid before eliminating them, rather than merely assuming that they probably were less valid because, for instance, they were elicited by the interviewer rather than being emitted sponteneously. Consequently, elimination rules tended to relate relatively closely to the substantive message contained in the claim, rather than to the circumstances under which the claim was provided. Nevertheless, recognizing that claims could vary in their degree of validity, two further categories of claims were eliminated as relatively less valid than the rest. First, references to the process of looking at or studying evidence were eliminated. 100 such references were found, including these:

- - District 9, Superintendent:

  I meet regularly with my four assistant superintendents. For any issue that comes up, we review and discuss both the interests and concerns as they exist in the system and the data.

- - District 115, Secondary Teacher:

  There are tests, written reports, oral reports, and homework, so hopefully the evaluation never rides totally on just one thing, or hopefully it reflects the whold child; it's not just that they've memorized something. [So you use a lot of different kinds of evidence to grade students?] Well, that's really a dream, but I like to strive for it.

Second, reasons why evidence had been sought or was considered valuable were eliminated. 395 reasons were found, including such comments as these:

- - District 7, Assistant Superintendent

  We're always trying to upgrade when it [the science curriculum] lists out. When our test scores and our judgment tell us things are not going as well as they could.

- - District 17, Program Director:

  These evaluation reports are helpful to familiarize you with what is going on, especially the narrative part.

These two types of statements were considered to be only marginally relevant to the actual use of evidence, and their removal reduced the data set to 814 citations, less than a third of the original set.

Table 1 summarizes the categories of citations that were eventually removed from the data base. Three comments should be made regarding the rationale for these rejection and retention rules. First, although the rejected citations are neatly categorized here, and accompanied by eligibility rules, no clear definitions of eligible or ineligible claims were available prior to sorting the citations themselves. These categories were developed in response to the citations available, using reasoning that was largely intuitive until the categories began to form. Second, the fact that a particular claim or type of claim is not considered to be a valid indicator of the use of evidence does not mean that the comment is altogether invalid. There is no a priori reason to believe, for instance, that the opinions expressed were not genuine opinions, that perceptions of how other people used evidence were always innacurate perceptions, or that the reasons people considered evidence to be valuable to them in their work were not real reasons. Indeed, these claims would constitute a useful data base for an investigation of what school district participants think about evaluation and test data in general. But the intent of this study was not to learn what people think about evidence, but rather to learn how evidence actually influences their thoughts and actions. Third, those citations which remain in the data base are still not all direct observations of use. Most still retain the status of indicators, in the sense that they are merely claims of use. However, they differ from those that were rejected in two important ways. First, they are more likely to be statements that were volunteered in the context of a substantive discussion, rather than responses to a direct question about how evidence was used, and thus they are analogous to those citations which Becker and his colleagues felt were more valid. Second, they are more likely to

Table 1

Number and Percent of Each Kind of Citation
Removed from the Data Base

Type of Citation	Number	Percent
REMOVED ON THE FIRST PASS THROUGH THE DATA		
Descriptions of technical procedures of data collection or dissemination	514	17
Uses of informal evidence such as observations or conversations	312	11
Opinions regarding use	608	21
Hearsay -- other people's uses	106	4
References to occasions when evidence was not used	126	4
REMOVED ON THE SECOND PASS THROUGH THE DATA		
References to the process of looking at or studying the evidence	100	3
Reasons why evidence is considered personally valuable	395	13
REMAINDER		
Valid indicators of the use of evidence	814	27

describe the specific substantive message contained in the evidence, rather than simply referring to a source of evidence as having been "helpful."

The difference between rejected and retained citations bears on the problem of false positives mentioned earlier. That the proportion of citations labeled "opinion" is almost as large as the proportion considered to be valid indicators of use (21 versus 27 percent) suggests that rhetoric about how evidence is or ought to be used can indeed present an obstacle to learning how evidence is in fact used. And many of the citations removed on the second pass through the data were false positives. These were citations in which school district participants said that they looked at or studied evidence, or stated why they found evidence to be valuable. They were not eliminated on the first pass through the data because the analysts inferred from these comments that the users really did what they said they did. But because participants' perceptions of their own behavior could be influenced by their opinions about what they should be doing, these comments were eventually rejected as relatively less valid indicators of use.

SORTING THE REMAINING DATA

Only 27 percent of the citations were finally considered to be valid indicators of use, with the percentage varying from 19 percent to 36 percent across school districts. These 814 citations were sorted according to the kind of uses they indicated. The categories of use that were eventually developed evolved primarily from the citations themselves, rather than from a priori categories, though the available literature suggested that certain types of uses would probably be found. Most of the definitions of use alluded to in the literature required some modifications in order to accomodate these citations.

One kind of use frequently discussed in the literature is conceptual use, or use for enlightment (e.g., Weiss, 1977). A respondent may say, for instance, that research helps her see things differently, or that it gives new insights into the social problem her program is intended to ammeliorate. Citations in this data base which indicated conceptual use included comments such as:

> - - District 50, Program Director:
> I find evaluation reports useful in stimulating me to think about the curriculum in new ways.
>
> - - District 57, Assistant Superintendent:
> Everything I work on, in some way or another there is data which I have read and I am considering consciously or unconsciously.

These citations suggest that evidence is indeed used to stimulate thought as well as to stimulate action. However, these citations were removed from the data base on the grounds that they constituted either opinions about the use of evidence or reasons why the respondent felt evidence was personally valuable. More direct indicators of conceptual use were found when respondents described what their new perceptions or insights actually were, and when the comments came up during discussion of substantive issues rather than in response to questions about how evidence is used. For instance, a Title I teacher who said she thought the Title I program in her district was a good one was asked why she thought so. Her response:

> - - District 50, Title I teacher:
> One way I can tell is that there's been a drop in Title I enrollment. It's gone from 380 to 270 to 186 in three years. Also I can look at the graphs on individual kids' progress. And kids are going through the readers faster than they used to. [There

> are five first-grade readers.] Four years ago, most of my students were in the third book at the end of the year. . . . Last year most of the kids were on grade level.

And a secondary school principal, in a conversation about curriculum development, brought up his staff's analysis of the test data, saying,

> - - District 4, Secondary Principal:
> We found that kids couldn't do without comprehension and we knew what we could do about it then.

Because these examples of conceptual use illustrate what people actually knew, learned, or had discovered, the category was labeled Personal Learning rather than Conceptual Uses, and it eventually contained two kinds of learning. One is the kind illustrated above, in which practitioners had drawn conclusions from the data. The other is purely descriptive knowledge, illustrated when interviewees said such things as:

> - - District 115, School Board President:
> We are losing 5,000 students a year. We lost 5,000 last year, we lost 5,000 this year, and we will lose 5,000 next year.
>
> - - District 7, Secondary Principal:
> Last year we went up in eleventh grade [test scores].

Thus although the sorting excersize began with the knowledge that a category labeled something like conceptual use might emerge, that label was abandoned in the face of the citations actually encountered. The citations themselves illustrate the results of the process of conceptual use, rather than the process itself, primarily because earlier retention rules required that descriptions of the process be eliminated. The citations that remain are more properly labeled as examples of personal learning.

Another kind of use suggested in the literature is forensic, or use for persuasion (e.g., Leviton and Hughes, 1981). Such uses have been referred to in previous literature, and examples exist in this data base as well.

- - District 27, Board Member:
> I convinced the board that we needed a management study done. . . . I knew they would never listen to me, but I thought they might listen to a study that came out of the business world. The study . . . made people aware that there were needs. I wasn't just a crazy lady saying it. It gave my position credibility.

- - District 50, Elementary Teacher:
> The psychologists are more willing to give tests to minority students if I can show them that they are not making progress.

But in addition to these citations were a number of others indicating that evidence accomplished several purposes in group interactions. For instance, it might be used to inform others

- - District 115, Observation of teacher's meeting:
> The team leader opened the meeting with a number of announcements. [including] "Next year we will have 123 fifth graders and 84 fourth graders."

or to respond to others

- - District 4, Board Member:
> A few years ago they [the administration] walked in here without a health curriculum, and much to everyone's surprise, the board said, "Hey, wait a minute." . . . We had the supporting data that showed that kids needed the services. They had been tested in health knowledge, sex education and nutrition.

and it could be used to supervise or to oversee the work of others as well

> - - District 19, Superintendent:
> Now I can ask questions of principals and department heads that we couldn't ask before. [Such as?] "What's going on in your freshman social studies course that causes the students to score less in your school than in the [XYZ] school?"

Thus, although the sorting began under the assumption that a likely category would be persuasion, the citations actually available suggested that a more inclusive category might be Interactions with Others, a category that could contain not only examples of persuasion, but also examples in which evidence was used to inform, supervise, or respond to others.

The category that presented most problems during sorting was one related to what had previously been labeled Instrumental Use (e.g., Rich, 1977; Leviton and Hughes, 1981). As in other cases, the category eventually used evolved over time, beginning with the notion that instrumental uses might constitute a category, and then developing in response to the citations actually available. One of the problems encountered was that the term instrumental use implies not only a use of evidence but a type of decision making as well. Instrumental uses are often assumed to entail major decisions about programs which are based on major summative evaluations of programs. Such uses were extremely rare in this set of 814 citations, although there were several examples in which evidence was used to make other kinds of decisions.

> - - District 72, Junior High School Principal:
> I looked at which teachers were responsible for which suspensions and referrals. I found out from my records that 70 percent of the problems of referrals and suspensions were caused by 17 percent of the teachers. So I decided to get rid of those problem teachers.

> - - District 35, Director of Physical Education:
> When students filled out evaluation forms for the dance class, they said they didn't like to have to take off their shoes and socks when they went into the gym. So I talked to the instructors about it and asked them to let the kids keep their shoes and socks on.

In addition to such decisions as these, there were a host of decisions about children which were based on test data. Children were grouped within classrooms, given course grades, and assigned to outside programs like Title I and special education. These decisions ranged from the fairly cut-and-dried procedures such as this

> - - District 220, Elementary Teacher:
> For the criterion tests students must get at least 80 percent mastery in order to move on, but there are always booster activities to re-teach the child. After that you can move them along. But if they don't pass the mastery test at the end of the book they can't go to the next book. If they fail the mastery test, then they have to go back over the whole book. Kids may be promoted to another class, but they will be using the same book.

to vague and judgmental procedures such as

> - - District 220, Deputy Superintendent:
> [Johnny Doe wanted special admission to the [ABC]school. Normally he wouldn't be eligible, but when I looked at his scores, I saw something funny, so I called his principal and . . .

Eventually, the category Instrumental Uses was replaced by the category Direct Applications, which in turn included three kinds of citations: those in which evidence was used to comply with an evaluation requirement; those in which evidence was used to sort or place students, including placements which relied on enrollment data to promote racial balance; and miscellaneous other direct applications. The miscellaneous group included such substantive applications as those illustrated above, as well as such diverse applications as defining criteria for program eligibility or racial imbalance, generating mailing labels, preparing grant applications, revising budget projections, and changing staffing distributions. These were combined into a single miscellaneous group not only because further subdivisions would have been very small, but also because discrete subdivisions would have been difficult to define. For instance, the director of a biology curriculum described the sequence of events he and his staff went through in developing a biology test:

> - - District 17, Program Director:
> I worked with the biology teachers and department heads to formulate the course objectives, and then we made a test and tested these objectives. We found that there were no changes resulting during the school year, that apparently the objectives were not being achieved. So then we changed the test but we still found no growth. I decided that the teachers were not really teaching these objectives. [Did you ever go out and observe the classes to see?] No I haven't done that, but experts have examined this test and the objectives and have approved of them. So if the test corresponds to the objectives and if no growth has been found on the test, teachers must be teaching something other than the objectives that they themselves set for biology. Next month I will bring them all together to find out what they really do teach. I think the objectives should be changed to match whatever is going on in the classrooms.

After much rumination and argument, this citation was finally considered to be an illustration of using tests to learn or to define what the curriculum actually is, rather than to define what it should be or to determine how well it is being taught. Several of the citations in the miscellaneous applications category illustrate similarly unusual or ambiguous uses of evidence.

Table 2 presents the proportion of citations that fall into each category of use, within each of four broad types of substantive issues that were discussed by participants in this study. Because sorting and placing children constituted such a large share of the citations that came up in conversations about classroom issues, two sets of proportions are presented in that column, one indicating the actual proportions and a second, in parentheses, indicating the distribution of citations that remain when that category is removed. Three caveates must accompany the interpretation of this table.

First, despite the fact that citations were sorted more than once and that ambiguous citations were disputed vigorously before being placed, the placement decisions are still highly judgmental. For instance, when a program director said,

> - - District 17, Program Director:
> I saw math scores go up after the change. Reading scores went down but are solid.

the citation could be considered to be personal learning of either a descriptive fact or an inference. The statement is mainly descriptive, and is not followed by a "therefore" statement, as many inferences were. However, implicit in the phrase, "after the change" is an inference that the changes in test scores were due to the change in the program. This citation was therefore categorized as an example of an inference. Similarly, when an elementary principal described his new curriculum-referenced testing system by saying,

Table 2

Percent of Citations falling into Each Category of Use

Type of Use	Context of Use				Weighted Average Percent of all Uses
	Policy Issues	Program Issues	Building Issues	Classroom Issues	
PERSONAL LEARNING					
Descriptive Knowledge	13%	10%	11%	5% (9%)	10%
Inferences, Conclusions	20	18	20	15 (30)	19
INTERACTIONS WITH OTHERS					
Informing	7	5	7	5 (10)	6
Persuading	18	19	10	6 (11)	14
Supervising	22	13	13	2 (4)	14
Responding	1	3	9	10 (19)	5
DIRECT APPLICATIONS					
Complying with Regulations	(.004)	1	2	3 (5)	2
Sorting and Placing Children	7	10	17	48 (--)	18
Miscellaneous Applications	11	20	10	6 (13)	12
TOTAL PERCENT	99	99	99	100 (101)	100

[1] Numbers in parentheses indicate the proportions of uses in each category after child placement decisions are removed from the count.

[2] Miscellaneous applications include such things as generating mailing labels, allocating staff, making minor modifications in curriculum or programs, determining the content of workshops, responding to a hotline telephone call from a parent, deciding what other research is needed, and so on.

> -- District 9, Elementary Principal:
> I know in an instant who is following the objectives and what children are doing and why.

he could be saying that this is a reason why he likes the system, or he could be saying that he actually uses it to supervise his teachers. Though the statement indicates a strong intent to use the system for supervision, it was classified here as no more than a statement of intent, and was eliminated from the final data base.

Second, these citations do not reflect a complete survey of all uses that occurred in these 16 school districts, or even all uses by the sampled respondents within these districts. Rather they are the uses that field investigators observed in the meetings they happened to attend and the uses that came up in interviews which were intentionally designed not to elicit all possible uses, but instead to discover a few uses embedded in their natural substantive context. Interviewers did, of course, steer the conversations toward topics they expected to yield more examples. If a program director said the two issues she was most perplexed about were complying with red tape requirements and upgrading the quality of the program, the interviewer would ask for elaboration of the second issue, on the assumption that that topic would be likely to entail more uses of evidence. But such gentle guidance is a far cry from asking for as many uses as can be generated within the next hour and a half.

A correlary to the second caveate is that the column frequencies shown in Table 2 are more likely to reflect the value of evidence for the issues that happen to be facing educational practitioners these days than to reflect any enduring inclinations on the part of the practitioners themselves. This fact is emphasized in the column headings of Table 2. Citations were sorted according to the type of issue discussed, rather than the type of person interviewed or observed. If a building principal discussed desegregation, his note was placed in the category of policy issues, and if a superintendent discussed placement of children in special education, her note was placed in the category of classroom issues. The patterns of use indicated in Table 2 should be inerpreted in this light. For instance, people who discuss program issues tended to mention more miscellaneous applications than others, and it is at the program management level, rather than at the level of building or district-wide management, that such activities as producing newsletters, providing in-service training, preparing grant applications, and so forth, tend to occur. This is not to say that the proportions shown in Table 2 indicate the precise distribution of occurrences of each kind of use, but rather that they indicate in a general way how evidence tends to get used in the resolution of contemporary educational issues.

The third caveate is that these data came originally from a sample of school districts nominated on the basis of how they used evaluation or test data. If these districts were indeed unusual or outstanding in their uses of evidence, these findings might not generalize. However, there are two reasons to believe that the data do generalize. One is that nominations were often found to be based on inaccurate second- or third-hand rumors. A district nominated because the evaluation unit was engaged primarily in policy research was found to have a unit that engaged mostly in testing, and did very little policy research. A district nominated because it used evidence in its annual reports to the state was found to invent most of the contents of its annual report. Although these sixteen districts are unusual in that they had been heard of by others, it is not clear that they differ from school districts in general in any other regard. The second reason for believing the findings generalize is that the findings regarding patterns of use are reasonable consistent across these 16 school districts, even though the districts vary in a number of other ways. In addition to their demographic variations described earlier, they differ in their organizational styles. Some were bogged down in political struggles or budgetary deficits that immobilized them, others hummed like machines. Some staffs critically analyzed every new idea, while other staffs embraced new ideas with the enthusiasm of cheerleaders, no questions asked. Some routinely involved parents in active debates over potential changes in practices while others hid facts from parents and tried to keep them out of decision-making processes. Yet despite these variations in style, only two districts had patterns that deviated substantially from the others. These districts both had unusually large proportions of opinions about the use of evidence. They had both instituted new management-by-objectives systems which relied heavily on test data, and members of these districts were divided in their opinions about the value of the systems. In both cases, the test data, or the management system in which they were embedded, constituted the substantive issue most often raised by interviewees, and consequently there were an unusually large proportion of opinions generated in these two districts. Aside from these opinions, however, the remaining citations from these districts are distributed roughly as those from other districts are, including the proportion of citations in which data were used for supervision.

These caveates suggest that the data should not be interpreted as indicating absolute frequencies of different kinds of uses of evidence. But they are nevertheless useful indicators of the dynamics of educational practice, and are informative in that sense. For instance, regardless of the issue being discussed, participants acquire more inferences and conclusions from the data than descriptive facts, a finding that suggests a tendency to think about the evidence rather than to simply retain it as factual knowledge. Second, of the three broad categories of use, interactions with others are more common than either of the other categories, thus suggesting that much of the work involved in settling issues and promoting change in education is done by groups rather than by individuals. Third, policy and program issues, whose resolutions tend to involve more people, tend also to generate more uses for persuasion than do building or classroom issues which are generally settled by individuals. Finally, references to big decisions are conspicuous by their near-absence, indicating either that such decisions really do not occur very often or that they are made by groups rather than by individuals and that perhaps for that reason are not mentioned by individuals in their own discussion of the issue. All of these interpretations of the data displayed in Table 2 are consistent with other research., with each other, and with other evidence gathered during the course of this study.

If the citations are taken out of Table 2, and returned to the notes from which they were taken, the interrelationships among individual inferences, group interactions, and piecemeal changes can be seen. For instance, an assistant superintendent in District 17 describes events related to spelling test scores in his district. It all began when he was perusing the annual printouts of test scores and noticed that spelling scores were lower than other subject areas (descriptive knowledge). He decided to commission a study of spelling instruction to determine why this might be the case (miscellaneous application). But to do so he had to get money from the school board (persuasion). Once he did that, the board decided to invest more money in spelling, both in new textbooks and in a massive in-service training program for teachers (hearsay application). Now the board is asking this assistant superintendent for evidence that spelling scores are going up (hearsay supervision). He now reports to them on a topic that they had never asked about before (informing). In fact, scores have not gone up and he told the board this is because the district is decentralized and schools are autonomous, and that several have decided not to adopt the new speller (inference, conclusion). He has asked the board to give him another year to get the scores up (responding) and he plans to spend the following year getting the schools to adopt the new speller (supervision).

And a secondary principal in district 18 describes this sequence of events. He studied absenteeism in his school and discovered that the greatest absentee rate occurred among tenth graders in the non-college-bound curriculum (descriptive knowledge). He discussed this finding with his faculty (informing) and they discussed the data and interpreted it (inference). They decided to make two changes in their building practices with regard to absenteeism (miscellaneous applications). The math teacher is now running the results through the computer to see whether these changes have been effective in reducing absenteeism. However, the principal has already decided to keep these new practices, even if they aren't effective in changing absenteeism, because:

> Parents like it, because they feel we are giving them feedback where we didn't before. Teachers like it because it gives them a feeling that at least something is being attempted, that the school is taking a stand. And even some of the students like it, because at bottom they respect the notion of accountability. Perhaps more important, it has gotten communication going between all the different groups (use of informal information).

The claims of use tallied in Table 2, then, are indicative of the dynamic processes which lead to educational changes. They indicate that such changes come about through series' of individual inferences, of interactions among individuals, and of minor adjustments, rather than from single meetings in which all evidence is reviewed and weighed so that a single decision can be made.

CONCLUSIONS

Qualitative research methods are often used to explore questions about subtle processes such as how evaluation and test data are used. But these methods present unique problems to educational researchers regarding how data should be collected and analyzed. This paper offers one approach to these problems. The approach is two-pronged. On one side is a data collection procedure which searches for uses that are embedded in substantive issues rather than asking self-consciously about use. On the other side is a method of identifying relevant, irrelevant, and marginally relevant claims about use.

The method has two disadvantages. First, the fact that it converts qualitative data into quantitative data may tempt readers to believe that the data are more precise than they really are and that they can be generalized using statistical procedures. But these numbers came originally from qualitative data and they carry with them all the inferential problems associated with haphazard sampling and all of the judgmental problems associated with the data collection procedures that characterize qualitative inquiries. Second, the procedure is extremely time consuming.

The apparent lack of inferential power and the labor intensiveness of the method raise questions as to its ultimate value. But the method offers several advantages other than simply enabling the analyst to summarize qualitative findings in a table such as Table 2. First, the distinctions resulting from this process do more than provide rules of evidence -- they clarify and define the nature of the phenomenon under study. Second, the tallies enable the analyst to check his impressions of prevalence against actual frequencies of various kinds of claims. Thus the analyst is less likely to be swayed by respondents' opinions or by his own predilections to believe some points of view more than others. Even though the figures themselves are tentative, they provide a means of testing impressions that would otherwise be even more tentative. Moreover, the numbers provide comfort to readers as well as analysts, particularly those readers who are unsure how reliable such statements as, "Most people said this," or "A few said that," really are. Most importantly, this analytic strategy encourages both the analyst and the reader to think in terms of the relative validity of different claims of use, rather than accepting every datum as face valid. Citations can be separated into those that are responses to questions versus those that are voluntarily emmitted, those that occur in interviews versus those that occur in group meetings, those that were gathered by one field investigator versus those gathered by another, those that describe general processes versus those giving specific examples, and so forth, thus facilitating clearer specification of the rules of evidence regarding what constitutes invalid claims as well as what constitutes valid claims. If educational researchers are to continue to rely on qualitative methods of inquiry, such critical review of evidence is essential.

NOTES

1. There were only two meetings which were denied to field investigators. Both of these were meetings of superintendents' cabinets during which personnel matters were to be discussed.

2. District code numbers indicate the number of students served by the district. District 4 serves approximately 4,000 students, and district 240 serves aroung 240,000 students. The code numbers vary randomly about the actual enrollment figures by ± 15 percent.

REFERENCES

Becker, H. S., Geer, B., Hughes, E. C., and Strauss, A. S. Boys in white: Student culture in medical school. Chicago: University of Chicago Press, 1961.

Leviton, L. and Hughes, F. X. Research on the utilization of evaluations. Evaluation Review, 1981, 5, 525 - 548.

Rich, R. F. Uses of social science information by federal bureaucrats: Knowledge for action versus knowledge for understanding. In C. H. Weiss (Ed.), Using social research in public policy making. Lexington, MA: D.C. Heath, 1977.

Weiss, C.H. Research for policy's sake: The enlightment function of social research. Policy Analysis, 1977, 3(4), 531 - 545.